J. Finan. Intermediation 53 (2023) 100996

Contents lists available at ScienceDirect

Journal of Financial Intermediation

journal homepage: www.elsevier.com/locate/jfi

Inefficient liquidity creation

Stephan Luck [a], Paul Schempp [b,*]

[a] Federal Reserve Bank of New York, United States of America
[b] University of Cologne, Cologne, Germany

ARTICLE INFO

JEL classification:
G21
G23
G28

Keywords:
Banking
Macroprudential regulation
Liquidity creation
Fire sales
Pecuniary externalities
Shadow banking
Regulatory arbitrage

ABSTRACT

We present a model in which intermediaries create liquidity by issuing safe debt. Two types of intermediaries emerge: Traditional banks that create liquidity by issuing equity and holding assets to maturity, and market-based intermediaries that create liquidity by selling assets in fire sales in downturns. We show that the reliance on market-based intermediation is necessarily too high, but liquidity creation is not. It can also be too low as the endogenous fire-sale risk can push liquidity creation below its optimum. We argue that standard capital or liquidity regulation are ineffective, and optimal macroprudential regulation should instead target market-based intermediation.

1. Introduction

A key economic function of financial intermediaries is to create liquidity by issuing safe debt (see, e.g., Diamond and Dybvig, 1983; Gorton and Pennacchi, 1990; Dang et al., 2015, 2017). However, the way financial intermediaries create liquidity has seen fundamental changes in recent decades. One particularly important phenomenon is the rise of market-based financial intermediation: The traditional banking model, in which insured deposits and equity are used to finance illiquid loans, has in part given in to a new, market-based form of financial intermediation. In this new regime, banks either rely on uninsured wholesale funding, or banking services are provided outside of the formal banking system by non-bank intermediaries, often referred to as shadow banks.

The empirical literature has made great progress in mapping out this new financial architecture (Gorton and Metrick, 2010; Cetorelli et al., 2012; Pozsar et al., 2013). Moreover, it is now well understood how market-based forms of bank funding and non-bank financial intermediation have contributed to the 2007–2009 financial crisis (Gorton, 2010; Financial Crisis Inquiry Commission, 2011; Gorton and Metrick, 2012; Acharya et al., 2013; Krishnamurthy et al., 2014). However, there is little guidance from economic theory on the optimal composition of the financial system and the optimal mix of traditional forms of banking on the one hand, and market-based financial intermediation on the other. This paper attempts to close this gap and asks two questions. First, how much liquidity should be created via market-based intermediation as opposed to non-market based hold-to-maturity banking? And, second, how is the financial system optimally regulated in the presence of these different types of financial intermediation?

We develop a model similar to the models suggested by Stein (2012) and Hanson et al. (2015) in which financial intermediaries' main function is to finance risky investments by issuing safe debt. When intermediaries issue safe debt, it is associated with a liquidity benefit. That is, by having the same payoff across all states of the world, safe debt generates utility for households beyond its guaranteed payoff.

Intermediaries can engineer safe debt and thus create liquidity in two ways. The first option is as in Stein (2012) where intermediaries can create liquidity by selling off risky assets in exchange for safe assets in a fire sale whenever an economic downturns looms, but before the worst state of the world is attained. While fire sales have the benefit of reducing uncertainty over intermediaries' asset values, investors that purchase assets forgo other profitable investment opportunities. Hence, fire sales entail private and social costs. The second option is to create liquidity by holding assets until maturity and issuing a sufficient amount of risk-absorbing equity claims (compare, e.g., Admati and Hellwig (2013), Hellwig (2015)).

* Corresponding author.
E-mail addresses: stephan.luck@ny.frb.org (S. Luck), pschempp@uni-koeln.de (P. Schempp).

https://doi.org/10.1016/j.jfi.2022.100996
Received 31 January 2022; Received in revised form 6 September 2022; Accepted 15 September 2022
Available online 3 October 2022
1042-9573/

In equilibrium, two types of bank business models coexist. On the one hand, a sector that resembles a traditional banking sector emerges, where intermediaries hold risky assets to maturity, and issue equity to absorb risks. On the other hand, a second sector of market-based banking emerges in which intermediaries create liquidity not by holding assets to maturity, but by selling assets in downturns. This sector can also be interpreted as a shadow banking sector.[1] In the former, intermediaries finance illiquid assets with deposits and equity, and are endogenously never subject to early withdrawals—despite the absence of government interventions such as deposit insurance. In the latter, intermediaries finance illiquid assets by issuing short-term debt instruments and are subject to withdrawals and fire sales in economic downturns. Importantly, both types of banking arise endogenously, and market-based intermediation in our setup does not necessarily need to be a response to regulatory requirements as is often explicitly or implicitly assumed in the theoretical literature on forms of market-based shadow banking (Plantin, 2015; Parlatore, 2016; Huang, 2018; Segura, 2018; Chrétien and Lyonnet, 2019; Martinez-Miera and Repullo, 2019). This is not to say regulatory arbitrage is not important. Indeed, regulatory arbitrage is a key driver of shadow banking activities (see, e.g., Buchak et al., 2018). Our model, however, suggests that market-based banking can also exist in absence of any privately costly regulatory requirements.

The key friction in our model is the intermediaries' inability to write state-contingent contracts with outside investors in the presence of aggregate risk (e.g., as in Holmström and Tirole 1998, 2011). This friction gives rise to a pecuniary externality which induces a wedge between the constrained-efficient allocation and the equilibrium outcome. As a consequence, too many intermediaries conduct market-based liquidity creation, and fire sales are excessive. Unlike in standard models with financial constraints and pecuniary externalities (Lorenzoni, 2008; Stein, 2012; Dávila and Korinek, 2018), the inefficiency does not result from excessive investment, but from too many intermediaries choosing a market-based business model.

An important consequence of the excessive extent of market-based intermediation is a general inefficiency of liquidity creation. Importantly, however, the direction in which liquidity creation is distorted in equilibrium is ambiguous: liquidity creation can be excessively high, but it can also be insufficiently low. In particular, in the latter case, the excessive reliance on market-based liquidity creation induces an endogenous fire-sale risk. If the wedge between the constrained efficient and the equilibrium extent of fire sales becomes sufficiently large, the aggregate level of safe debt created can be pushed below its optimal level.

This finding is thus in contrast to the conclusions in Stein (2012), who finds that liquidity creation is necessarily too high. This difference is due to the fact in our model, we allow the lowest possible payoff across states to be positive and thus there is scope for hold-to-maturity banking next to market-based banking. This, in turn, induces a non-monotonic relationship between aggregate fire sales and aggregate safe-debt creation.

The underlying mechanism can be illustrated as follows: when intermediaries sell their risky assets in exchange for safe assets upon the arrival of adverse news, they sell them at a price that is higher than the lowest possible cash flow. Thereby, ex-interim fire sales can, in principle, allow intermediaries to create a higher level of safe debt ex-ante. However, the extent to which liquidity can be created by fire sales does not only depend on the quantity of risky assets sold, but also on the price at which they are exchanged against safe claims. Fire-sale pricing may be such that any additional sales of a unit of risky assets increases the overall revenue by less than the lowest possible cash flow that the marginal unit sold would have generated on the balance sheet. If this is the case, the endogenous fire-sale risk reduces aggregate liquidity creation at the margin instead of increasing it. We show that such insufficiently low levels of liquidity creation are more likely to be found in economies in which fire-sales are unlikely to begin with, in which liquidity benefits from safe debt are high, and when secondary asset markets are illiquid.

Given the general inefficiency of liquidity creation, important implications for how to optimally regulate financial intermediation follow. First, we show that standard tools of banking regulation such as macroprudential capital or liquidity requirements may not be effective tools to address the excessive reliance on market-based intermediation and asset sales. If liquidity creation is insufficient in equilibrium, leverage is not excessive to begin with, and the optimal minimum capital requirement or liquidity regulation cannot be binding. Hence, in contrast to existing models with pecuniary externalities (Stein, 2012; Bianchi and Mendoza, 2018; Walther, 2016; Gersbach and Rochet, 2012, 2017; Malherbe, 2020; Kara and Ozsoy, 2020), our model indicates that capital requirements and liquidity regulation are no panacea when fire sales are excessive as they do not necessarily target the underlying problem of excessive market-based intermediation.

Second, we show that optimal macroprudential regulation should address the excessive reliance on market-based liquidity creation. This can either be done by directly restricting the extent of market-based intermediation, or by imposing a tax on asset sales. Finally, we show that if regulatory arbitrage is possible and restrictions or taxes can be avoided by shifting activities to an unregulated shadow banking sector, the above policy measures are rendered ineffective. However, the regulator can nonetheless implement the constrained-efficient allocation by subsidizing hold-to-maturity banking rather than restricting or taxing market-based intermediation (similar to, e.g., Ordoñez, 2018).

2. Setup

The economy goes through a sequence of three dates, $t = 0, 1, 2$. There is one good that can be used for investment and consumption. The economy is populated by three types of agents: households, intermediaries, and outside investors. For simplicity, assume that agents do not discount future consumption. In the following, we describe the agents' preferences, endowments, and technologies.

Households. There is a unit mass of households that are endowed with one unit of the good at $t = 0$. They maximize their expected lifetime consumption, given by $U = c_0 + E[c_1 + c_2] + \gamma \cdot d$, where d denotes the face value of risk-free debt a household holds at $t = 0$. A safe debt claim is defined as a claim that has the same payoff across all states of the world. The term $\gamma \cdot d$ denotes the liquidity benefits that households derive from holding safe claims. Thus, a household's required expected return on risky assets is given by 1, whereas the required gross interest rate on risk-free debt claims is given by $r = 1/(1+\gamma) < 1$. For simplicity, it is assumed that households cannot invest in the economy's technologies directly, but they can only hold claims issued by intermediaries that invest on behalf of the households.

Intermediaries. There is a unit mass of intermediaries who are the protagonists of the model. Their main role is to intermediate between households and the production technology, thereby creating liquidity through issuing safe debt. As specified in more detail below, intermediaries choose how much liquidity to create via holding assets to maturity and how much via asset sales.

Intermediaries are risk neutral, maximize their expected profits, and have full bargaining power vis-à-vis households. Intermediaries have no goods endowment but have access to a production technology.

[1] Our model abstracts from the fairly complex intermediation chains involved in the market-based banking in which some institutions originate and distribute assets and others hold them. We focus on capturing the fact that, on the liability side, traditional banking relies on stable deposits, while market-based banking (such as shadow banking but also market-based banking within regulated banks) tends to effectively rely relatively more on run-prone wholesale funding.

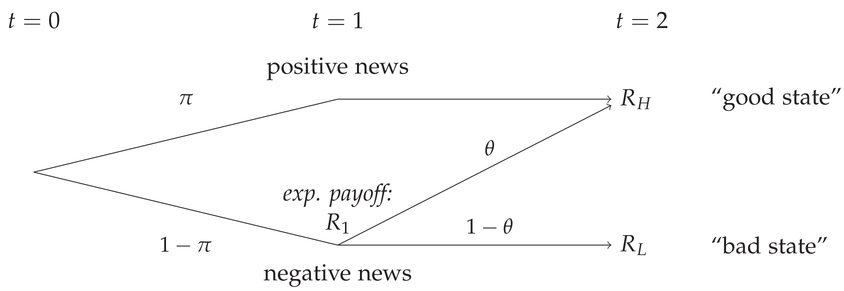

Fig. 1. Timing, Returns, and Information structure.

The technology should be interpreted as a shortcut to modeling the intermediaries' lending activity in the loan market. It is assumed that it has a fixed scale, requiring an investment of one unit.

The return of the technology R_s accrues at date 2 and is subject to aggregate risk. The macro state of our economy can either be high, $s = H$, or low, $s = L$, with $R_H > R_L$. Whereas the state realizes at date $t = 2$, news about the state of the world arrives in the form of a public signal at date $t = 1$, see Fig. 1. Positive news arrives with probability π, indicating that the high state H will realize with certainty, so the return is $R_H > 1$ and risk-free. With probability $1-\pi$, there are negative news, and given this signal, the expected return is R_1: Given negative news, the conditional probability of state H is given by θ, and that of state L is $1-\theta$, implying that $R_1 = \theta R_H + (1-\theta)R_L$. The arrival of negative news can be interpreted as news that the probability of a recession increases. But crucial to our model is the property of this payoff and information structure that there remains uncertainty about returns after negative news, and that the lowest possible realization after negative news is lower than after positive news. The above assumptions imply that the ex-ante probability of $s = H$ is given by $\pi + (1-\pi)\theta$, and $(1-\pi)(1-\theta)$ for $s = L$. In the later analysis, whenever we are interested in the effect of a change in R_L, we will implicitly also vary θ in order to keep R_1 constant. This way, a reduction of R_L will constitute a mean-preserving spread of the asset's payoff.

Remark (*Crucial Assumptions of Payoff Structure*)**.** The assumed two-state economy is the simplest version. For the model mechanics, there are two important assumptions: First, the lowest possible cash-flow after negative news must be smaller than the lowest possible cash-flow after positive news, and second, cash-flow must remain risky after negative news.

Given that intermediaries have no goods endowment, they need to raise funds from households. The model distinguishes between two types of funding: safe debt and risky equity claims. Because the gross interest rate on safe debt is given by $r = 1/(1+\gamma)$, each intermediary can raise an amount $d/r = (1+\gamma)d$ through issuing safe debt with a face value d. Given that each intermediary has a fixed size and has to raise one unit of financing, the remaining funding is raised through equity issuance yielding $e = 1 - (1+\gamma)d$.[2]

In order to create liquidity by generating safe debt claims, an intermediary must ensure that the cash flow in the low state, $s = L$, amounts to at least d. Thus, intermediaries are subject to a financial constraint which we refer to as a "safe-debt constraint".[3] There are two ways of engineering safe debt claims. Either an intermediary holds assets to maturity, allowing her to generate an amount of safe debt equivalent to the lowest possible cash-flow of her technology, R_L. Alternatively, intermediaries can sell assets to outsiders after the arrival of negative news at $t = 1$ and thus before the final state (possibly $s = L$) has realized, given that the price higher than R_L. We refer to the former as *hold-to-maturity* banking—resembling a traditional bank business model—and the latter as *market-based* banking. Note that the parameter R_L naturally captures the relative attractiveness of hold-to-maturity banking over market-based banking, with the former being more attractive for higher R_L.

As indicated above, our model abstracts from the fairly complex intermediation chains involved in the market-based banking in which some institutions originate and distribute assets and others hold them. In contrast, we focus on capturing the fact that, on the liability side, traditional banking relies on stable deposits, while market-based banking tends to effectively rely relatively more on run-prone wholesale funding.

Investors. As indicated, assets can be bought by outside investors. A unit mass of such outside investors is born at $t = 1$. Outside investors can either be interpreted as intermediaries that are not subject to macroeconomic risk, or different types of financial institutions that are unconstrained and able to purchase bank assets in downturns such as financially other intermediaries, deep-pocketed investors, investment funds or insurance companies.

Outside investors are risk neutral and maximize their date-2 payoff. Each investor is endowed with W units of the good and has access to a production technology. The production technology of investors is specified as follows: An investment of k at $t = 1$ yields an output of $g(k)$ at $t = 2$ with certainty, where the function $g(\cdot)$ is twice continuously differentiable and strictly concave with $g'(W) \geq 1$ and $g'(0) \to \infty$. Hence, investors need to decide how much of their endowment they use to operate their technology, and how much they use to purchase assets from intermediaries.

Note that the key friction of the model is that outside investors are not born before $t = 1$, and hence they cannot contract with intermediaries in $t = 0$, which is equivalent to a limited-commitment friction as in Holmström and Tirole (1998). This assumption essentially means that in the initial period, *intermediaries cannot contract their future financing conditions* and thus cannot insure against the occurrence of positive or negative news. In the absence of this friction, state-contingent contracts available at $t = 0$ would allow the implementation the first-best allocation (see Appendix B). However, we will show that in the presence of this friction, the laissez-faire equilibrium does not coincide with the constrained-efficient allocation.

3. Equilibrium analysis

We start out with describing how intermediaries can create liquidity, then characterize the asset-market equilibrium at date 1, before finally turning towards the analysis of the laissez-faire equilibrium and the constrained-efficient allocation.

[2] Throughout our main analysis, we do not discuss the maturity of intermediaries' debt claims. In Appendix C, we show under which conditions the optimal debt contract is short-term and how this can give rise to runs. Moreover, the "equity" could also comprise junior loss-absorbing financing instruments.

[3] We prefer the term "safe-debt constraint" over "collateral constraint" because the assets are not actually encumbered and pledged to lenders. Second, as described above, collateral constraints typically depend only on current market prices instead of possible future realizations of cash-flows (compare, e.g., Dávila and Korinek, 2018).

3.1. Liquidity creation

As described above, intermediaries can create liquidity in two different ways: They can hold assets until maturity, or they can sell them to outsiders upon the arrival of negative news. Formally, let μ_i denote the amount of assets that intermediary i sells; it thus represents the reliance on market-based liquidity creation. Correspondingly, $1-\mu_i$ describes the degree of hold-to-maturity banking. After negative news, the expected value of intermediaries' assets is R_1, but intermediaries can only sell them to outside investors at a price δR_1. The fire-sale discount factor $\delta \leq 1$ is a key endogenous variable in our model and will be determined below.

The amount of safe debt intermediary i creates is then given by

$$d_i(\mu_i) = \underbrace{(1-\mu_i) \times R_L}_{\text{hold-to-maturity}} + \underbrace{\mu_i \times \delta R_1}_{\text{market-based}} . \tag{1}$$

The amount of safe debt created per unit of assets that is held to maturity is R_L, whereas the amount of safe debt per unit of assets sold is given by δR_1 and thus depends on the fire-sale discount δ.[4]

The profit of intermediary i as a function of the amount of assets sold, μ_i, is given by

$$\Pi_i(\mu_i) = \pi R_H + (1-\pi)\left[(1-\mu_i)R_1 + \mu_i \delta R_1\right] - 1 + \gamma \underbrace{\left[(1-\mu_i)R_L + \mu_i \delta R_1\right]}_{d_i} . \tag{2}$$

The amount of assets being sold in the secondary market after negative news is given by $\mu = \int \mu_i di$, and it can be interpreted as the intermediation sector's aggregate reliance on the market-based business model.

3.2. Asset-market equilibrium

In the next step, we derive the equilibrium of the asset market at $t = 1$, in which a mass μ of assets is being sold by intermediaries to outside investors when negative news arrive.[5]

Let M denote the amount of funds used by outside investors to buy assets when negative news arrives and the investment in the late technology is thus given by $k = W - M$. M is a crucial variable in this model as it describes the amount of liquid funds that outside investors devote to purchasing assets from intermediaries in a fire sale instead of operating their productive technology. Increasing M has two opposing effects on welfare: On the one hand, it reduces profitable late investment after the arrival of negative news. On the other hand, it increases the intermediaries' scope to generate safe debt and liquidity benefits ex-ante because it increases their cash flow after negative news.

Given a fire-sale discount factor δ, expected cash-flow of outside investors at date 2 as a function of M is then given by $g(W-M) + \frac{M}{\delta}$. Their optimal portfolio decision M for a given discount δ is thus characterized by

$$\delta = \frac{1}{g'(W-M)} \tag{3}$$

as long as investors are indifferent at the margin and thus choose an interior value $M \in (0, W)$.[6] Thus, assets are priced according to the marginal productivity of outside investors as in, e.g., Lorenzoni (2008) and Stein (2012).

Given that a mass μ of asset are being sold by intermediaries after negative news, market clearing is characterized by

$$M = \mu \delta R_1. \tag{4}$$

Combined with the outside investors' indifference condition (3), we obtain the condition for an asset-market equilibrium:

$$\mu(M) = \frac{M g'(W-M)}{R_1}. \tag{5}$$

This equation characterizes the unique mapping between the amount of funds used to purchase assets, M, and the degree of market-based banking, which is equal to the amount of assets sold, μ.

Let us define $\overline{M}$ as the largest amount of funds that can be used to purchase assets in a fire sale. It is given by the amount of funds used to purchase assets in case all intermediaries choose to sell all their assets, i.e., $\mu(\overline{M}) = 1$.

3.3. Excessive fire sales in the laissez-faire equilibrium

Given the characterization of the asset-market equilibrium after negative news at date 1, the laissez-faire equilibrium of the economy can be characterized as follows. An equilibrium of the economy is given by a pair (μ^*, M^*) that satisfies the asset-market equilibrium condition (5), where $\mu^* = \int \mu_i^* di$ and the μ_i^*s maximize the intermediaries' profits (2).

Note that in our model, the linearity of individual liquidity creation (1) and profits and (2) in μ_i implies the distribution of individual choices of μ_i does not matter as long as it leads to the same aggregate $\mu^* = \int \mu_i di$. If hold-to-maturity and market-based banking are both used in equilibrium, then each intermediary must be indifferent between the two business models. Therefore, each intermediary could choose the same mixed business model $\mu_i = \mu^*$, but this outcome is equivalent to a situation where a fraction μ^* of intermediaries chooses $\mu_i = 1$ (market-based intermediaries), and the remaining intermediaries choose $\mu_i = 0$ (hold-to-maturity intermediaries). When explaining our results in the following, we will adopt the interpretation that intermediaries choose separated business models. Note that we provide a modification of the original model in Appendix D in which the strict separation of intermediaries' business models is strictly optimal.

The profit functions of hold-to-maturity intermediaries (HTM) and of market-based intermediaries (MB) are given by

$$\Pi_{HTM} = \pi R_H + (1-\pi)R_1 - 1 + \gamma R_L, \quad \text{and} \tag{6}$$

$$\Pi_{MB}(\delta) = \pi R_H + (1-\pi)\delta R_1 - 1 + \gamma \delta R_1. \tag{7}$$

Let us define $\tilde{\delta}$ as the fire-sale discount that equates the profit from market-based and hold-to-maturity banking,

$$\tilde{\delta} := \frac{(1-\pi)R_1 + \gamma R_L}{(1-\pi+\gamma)R_1}. \tag{8}$$

In the following, we will first analyze interior solutions only (i.e., allocations in which outside investors use some but not all of their funds, and some but not all intermediaries conduct market-based banking). Such interior solutions occur under the following technical assumption:

Assumption 1.

$g'(W) < \tilde{\delta}^{-1} < g'(W - \overline{M})$.

The first part of this inequality implies that if investors did not purchase assets, $M = 0$, their marginal productivity would be so low and thus the fire sale discount factor so high that all intermediaries would want to become market-based intermediaries, which is inconsistent with no asset sales. Thus, we will always have $M > 0$. The second part implies that even if all intermediaries sold their assets (and investors would thus use the maximum level $\overline{M}$ for asset purchase), investors' marginal productivity would be so high and thus the fire sale discount

[4] Note that we implicitly assume that intermediaries cannot securitize their assets. Securitization would allow intermediaries to engineer safe and risky tranches, with only the latter being sold after negative news. However, our results are qualitatively robust to allowing for securitization as long as an intermediary cannot securitize her *entire* balance sheet.

[5] Assets are never traded after the arrival of positive news.

[6] If $g'(W) > \delta^{-1}$, the investors would choose the corner solution $M = 0$. The corner solution of $M = W$ is ruled out by the assumption of $g(0) = \infty$. Thus, outside investors use some part $M \in (0, W)$ of their endowment to purchase assets whenever the equilibrium discount satisfies $g'(W) < \delta^{-1}$.

factor so low that no one would want to engage in market-based intermediaries, which is inconsistent with selling the entire portfolio. Thus, we will always have $M < \overline{M}$.

The equilibrium is then characterized as follows:

Lemma 1. *When Assumption 1 is satisfied, the laissez-faire equilibrium is characterized by* $M^* \in (0, \overline{M})$ *such that* $g'(W - M^*) = \tilde{\delta}^{-1}$, *and* $\mu^* = \frac{M^* g'(W-M^*)}{R_1} \in (0,1)$.

Proof. See Appendix.

In equilibrium, the fire-sale price must be such that intermediaries are indifferent between market-based banking and hold-to-maturity banking. For instance, assume that $\mu < \mu^*$. In this case, the profit an intermediary makes by conducting market-based banking is higher than the profit from conducting hold-to-maturity banking. As a consequence, some intermediary would prefer to migrate her business into the market-based sector. To attain the equilibrium, the number of intermediaries conducting market-based liquidity creation must thus increase such that the fire-sale discount factor is $\delta^* = \tilde{\delta}$ and expected profits are the same across both sectors. Thus, the model features the coexistence of traditional hold-to-maturity banking and market-based financial intermediation in absence of regulatory intervention (see also Ari et al., 2017). This is in contrast to most existing models of market-based shadow banking that derive the existence of shadow banks as a reaction to regulatory interventions.

Note that the leverage of market-based banks is strictly larger than that of hold-to-maturity banks: While the lowest asset value is R_L for hold-to-maturity intermediaries, it is $\delta R_1 > R_L$ for market-based ones. Thus, the leverage choice reveals an intermediary's chosen business model already at date 0.

Having characterized the laissez-faire allocation, we now turn to the welfare analysis. Welfare can be characterized as follows:

$$\mathcal{W}(\mu, M) = \gamma \underbrace{[(1-\mu)R_L + M]}_{D=\int d_i di} + (1-\pi)[g(W-M)+M] + \pi g(W) + (1-\pi)R_1 + \pi R_H. \tag{9}$$

The first term denotes the liquidity benefit, and the second term denotes the gross return of outside investors' endowment after the arrival of negative news, multiplied with the probability of negative news occurring. Note that the funds used for asset purchase, M, are not lost, they just get transferred to intermediaries. The second line denotes constant parts of the welfare function: the investors' payoff after positive news and the expected payoff of intermediaries' assets. The two terms in the first line capture the social planner's tradeoff: Creating more liquidity benefits ex-ante by creating more aggregate safe debt D reduces efficient late investment ex-interim after negative news.

The constrained-efficient allocation (μ^{**}, M^{**}) maximizes welfare (9) subject to the asset market clearing condition (5).

Proposition 1 (*Excessive Fire Sales*)**.** *When Assumption 1 is satisfied, the laissez-faire equilibrium features strictly excessive market-based liquidity creation and excessive fire-sales compared to the constrained-efficient allocation:* $\mu^* > \mu^{**}$, *and* $M^* > M^{**}$.

Proof. See Appendix.

To derive the constrained-efficient allocation, we have to consider how μ affects welfare, taking into account how μ affects the amount of funds used to purchase assets, M, according to the asset-market equilibrium function (5). The marginal productivity of late investors in the constrained-efficient allocation can be expressed in terms of $g'(W - M^*) = \tilde{\delta}^{-1}$ (which characterized the laissez-faire equilibrium), and a term that represents the negative pecuniary externality of asset sales:

$$g'(W - M^{**}) = \tilde{\delta}^{-1} + \underbrace{\frac{\gamma R_L M^{**} g''(W - M^{**})}{(1-\pi)R_1 + \gamma R_L}}_{<0,\ \text{pecuniary externality}}. \tag{10}$$

Thus, the marginal productivity of outside investors is strictly smaller in the constrained-efficient allocation than in the laissez-faire equilibrium, implying that too much funds are used to purchase assets. I.e., from the social planners viewpoint, too many assets are being sold and too much liquidity is being created via market-based banking.

As in models with collateral constraints that feature excessive fire sales, intermediaries do not internalize that by conducting market-based liquidity creation and selling assets after the arrival of negative news, they affect the investment decisions of outside investors and thus the fire-sale discount. Thus, in equilibrium, too many assets are being sold and there is too little investment in the late technology. However, while excessive fire sales are typically linked to excessive investment ex-ante in standard models of fire sales and collateral constraints (see, e.g., Lorenzoni, 2008; Stein, 2012; Dávila and Korinek, 2018), the pecuniary externality induces an inefficient composition of the financial system in our setup.

3.4. Inefficient liquidity creation

So far, we established that the reliance on market-based liquidity creation and the extent of fire sales is necessarily excessive. We now show that the aggregate amount of liquidity that is created in equilibrium can nonetheless be both, too high and too low compared to the constrained-efficient allocation. This results from the fact that the amount of aggregate fire sales in the laissez-faire equilibrium can become so large that the marginal aggregate liquidity production becomes *negative*, i.e., aggregate liquidity would actually be larger with less reliance on market-based banking. The endogenous fire sale risk can thus push the amount of safe debt created below its optimal level.

Proposition 2 (*Inefficient Liquidity Creation*)**.** *The liquidity created by the amount of safe debt in the laissez-faire equilibrium can be lower than in the constrained-efficient allocation,* $D(\mu^{**}) < D(\mu^*)$. *Such inefficiently low levels of liquidity creation can only occur for intermediate values* $R_L \in (\underline{R_L}, \overline{R_L})$. *It holds that* $\overline{R_L} < R_1$, *and if for* $R_L = 0$ *it holds that* $\mu^{**} = \mu^* = 1$, *then it also holds that* $\underline{R_L} > 0$.

Proof. See Appendix.

Proposition 2 contains two statements. First, despite the fact that fire sales are always excessively high, liquidity creation is not always excessively high, but it can also be insufficiently low. Second, insufficiently low levels of liquidity creation are most likely for intermediate values of R_L, i.e., when traditional hold-to-maturity banking is neither too attractive nor too unattractive.

To build intuition for the first statement in Proposition 2 and why liquidity creation can be sufficiently low, it is helpful to compare the private incentives to conduct market-based liquidity creation with the social cost and benefits of asset sales. An individual intermediary i takes the discount δ as given when choosing μ_i and the individual liquidity creation is given by (1). In contrast, the aggregate level of liquidity created is given by the payoff from the total assets held until maturity, plus the funds raised from outside investors through selling risky asset at $t = 1$, i.e.,

$$D(\mu) = \underbrace{(1-\mu)R_L}_{\text{hold-to-maturity}} + \underbrace{M(\mu)}_{\text{market-based}}, \tag{11}$$

where $M(\mu)$ is the inverse of (5), the function that describes the asset-market equilibrium.[7]

A single intermediary that decides to conduct market-based banking has the following marginal effect on aggregate liquidity: On the one hand, hold-to-maturity liquidity creation declines by R_L. On the other hand, the aggregate amount of market-based liquidity creation increases by $M'(\mu)$. Note that $M(\mu)$ is strictly increasing, but concave because more reliance on market-based banking leads to lower a fire-sale price. Hence, the marginal amount of market-based liquidity created is decreasing in the extent of market-based liquidity creation. Given that the marginal liquidity created from holding assets to maturity is a constant and given by R_L, the total marginal effect of asset sales on aggregate liquidity creation can thus in principle become negative and $D'(\mu) < 0$.

Intuitively, for liquidity creation to be excessive, the wedge between the amount of assets sold in the laissez faire and the constrained efficient needs to be sufficiently large. Observe that the social planner never chooses a degree of fire sales that would destroy liquidity at the margin. After all, if the marginal asset sold were to reduce the overall liquidity created, the social planner could always increase welfare by reducing fire sales, increasing both outsider investor returns and liquidity creation. Thus, as can be seen in the numerical examples provided in Fig. 2, $D'(\mu^{**}) > 0$. Hence, a small deviation from the social planner's preferred constrained-efficient degree of fire sale—increasing fire sales marginally—can only create more and not less liquidity. Insufficient liquidity creation, in contrast, can occur when the wedge in assets sales and the reliance on market-based liquidity creation is sufficiently large such that the fire sale price in the laissez faire is substantially lower than in the constrained efficient.

The left panel of Fig. 2 illustrates a case in which the wedge is relatively small and liquidity creation is excessive., i.e., $D(\mu^*) > D(\mu^{**})$. In contrast, the right panel of Fig. 2 provides a numerical example in which the wedge is much wider and it indeed holds that $D'(\mu^*) < 0$. In such a case, liquidity is destroyed at the margin in the laissez-faire equilibrium: Any marginal intermediary, by deciding to perform market-based intermediation and thus selling one additional unit of assets, decreases the overall liquidity generated by the entire banking sector. Moreover, in this specific numerical example, liquidity is not only reduced at the margin—a necessary condition for insufficient liquidity creation—but the level is indeed also lower than in the constrained-efficient.

Said differently, fire-sale pricing may be such that any additional sales of a unit of risky assets increases the overall revenue by less than the lowest possible cash flow that the marginal unit sold would have generated on the balance sheet. If this is the case, asset sales reduce aggregate liquidity creation at the margin instead of increasing it.

Next, we discuss why liquidity creation can only be insufficiently low if R_L, the cash flow for intermediaries in the bad state, is neither too high nor too low. Consider first the case in which R_L is high. A high R_L implies that the potential to create liquidity with hold-to-maturity banking is very high and there is little potential to create liquidity by selling these assets at a discount. If R_L is sufficiently high, there is no wedge and it is socially *and* privately optimal not to sell any assets and, all liquidity is created via hold-to-maturity banking. Starting at such a very high level of R_L, consider the effect of decreasing R_L. Eventually a point is reached where assets are sold both in the constrained-efficient and in the laissez-faire allocation. However, assets sales in both constrained-efficient and laissez faire are a continuous function of R_L, implying that the wedge is small initially (as R_L is lowered).

Similarly, consider the case when R_L is low. A low R_L implies that the potential to create liquidity with hold-to-maturity banking is very low and thus assets sales are high in both the constrained-efficient and the laissez-faire allocation. In particular, for very low R_L, the wedge entirely disappears as it becomes privately *and* socially optimal to sell all assets. When increasing R_L from a very low level, eventually a critical threshold is reached, and it becomes optimal to conduct some hold-to-maturity banking in the laissez-faire allocation. The initial wedge, however, is again small due to the continuity of asset sales in R_L. This implies that an interval in R_L for which there is insufficient liquidity creation must be framed by intervals of excessive liquidity creation on both sides.

Finally, note that a crucial ingredient for the existence of insufficient liquidity creation is that $R_L > 0$. In contrast, if the banks' assets did not generate any cash-flow in the low state, $R_L = 0$, it would hold that $D'(\mu) \geq 0$ globally, so a potentially higher amount of asset sales in the laissez-faire would necessarily lead to excessive liquidity. This is the crucial difference to the model proposed by Stein (2012).[8] We believe its plausible to assume that $R_L > 0$ when intermediaries are able diversify their loan portfolio to some degree. Thus, a bank will never find itself in a situation in which its total asset portfolio generate a cash-flow of zero. Hence, our findings imply that a model that is conceptually similar to models with a collateral externality does not necessarily feature excessive liquidity creation or excessive leverage as suggested in Stein (2012).[9] Although fire sales are always excessive in our model as in the standard models with collateral externalities (Dávila and Korinek, 2018), liquidity creation can also be too low as the endogenous fire sale risk can push the level of safe debt below its optimal level.

3.5. Comparative statics

We next analyze how other parameters of the model affect the wedge in asset sales and liquidity creation and ask: in which types of economies is insufficient liquidity creation more likely? In particular, we analyze the effect of π (the probability of receiving good news and thus not having to sell any assets), of γ (liquidity benefit), as well as the effect of the shape of late investors' production function.

3.5.1. Comparative statics with respect to π

Consider first the effect of π, the probability of receiving good news and thus avoiding the fire sale. The left panel of Fig. 3 plots the regions in which the difference in the liquidity generated in the equilibrium ($D(\mu^*)$) and in the constrained efficient ($D(\mu^{**})$) is positive/negative/zero for different levels of R_L and π. We allow for all feasible values of R_L, i.e., $R_L \in [0, R_1]$ and choose π such that the likelihood of bad news ranges from very likely (if time was measured in year, $\pi = 0.9$ implies that a fire sale would take place in one out of ten years) to very unlikely ($\pi = 0.99$ and a crisis in one out of every hundred years). The left panel shows that insufficient liquidity creation (the blue shaded area) is more common in economies with a higher

[7] This inverse exists and is well defined because $\mu(M)$ is continuous and strictly monotone; the asset market equilibrium condition implies that more assets sold correspond to a larger budget used to purchase them.

[8] Assuming $R_L = 0$ automatically implies that assets which remain on the balance sheet after bad news has arrived do not generate any liquidity benefits. Therefore, increasing asset sales could never reduce the amount of liquidity created. This assumption also implies that in Stein (2012), there is no difference between D and M—the amount of liquidity creation D ("money") is exactly equal to the amount of funds that are used to purchase assets M.

[9] Concerning this issue, Stein (2012) writes: "In a more general model where the lowest possible value of output at time 2 is greater than zero, banks can issue some riskless long term debt [...] banks will continue to be tempted to issue short-term debt in this more general version of the model, and all the qualitative results that follow will continue to apply". [p. 64]. In contrast, our analysis shows that the qualitative results can be different and the amount of liquidity created can be too low.

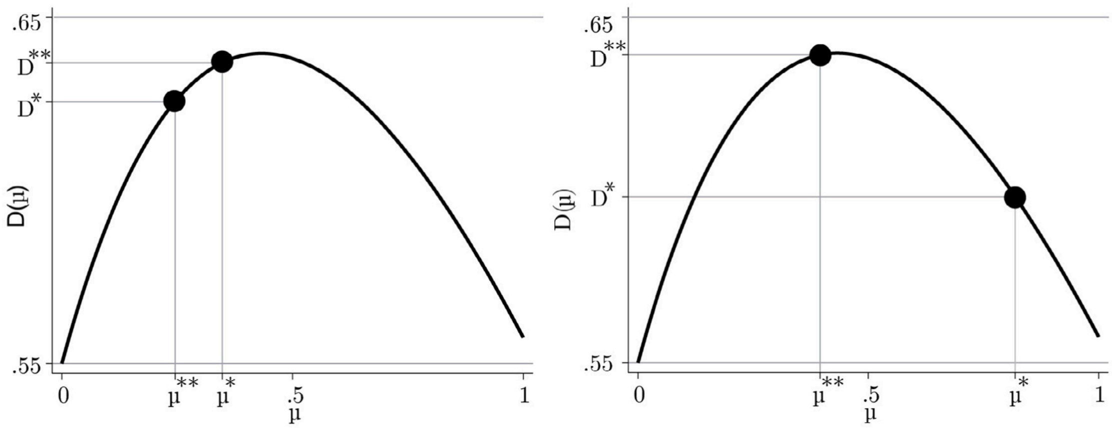

Fig. 2. This figure plots the aggregate liquidity production function $D(\mu) = (1-\mu)R_L + M(\mu)$ as well as the laissez-faire and constrained-efficient allocations for the following parametrization: $g(k) = \alpha ln(k)$ with $\alpha = 1.21$, $\gamma = 0.15$, $R_H = 1.2$, $W = 1.2$ and $R_1 = 1.05$. In the left panel ($\pi = 0.85$), there is excessive liquidity creation, whereas in the right panel ($\pi = 0.98$), there is not only liquidity destruction in the laissez-faire, but even insufficient liquidity creation compared with the constrained-efficient.

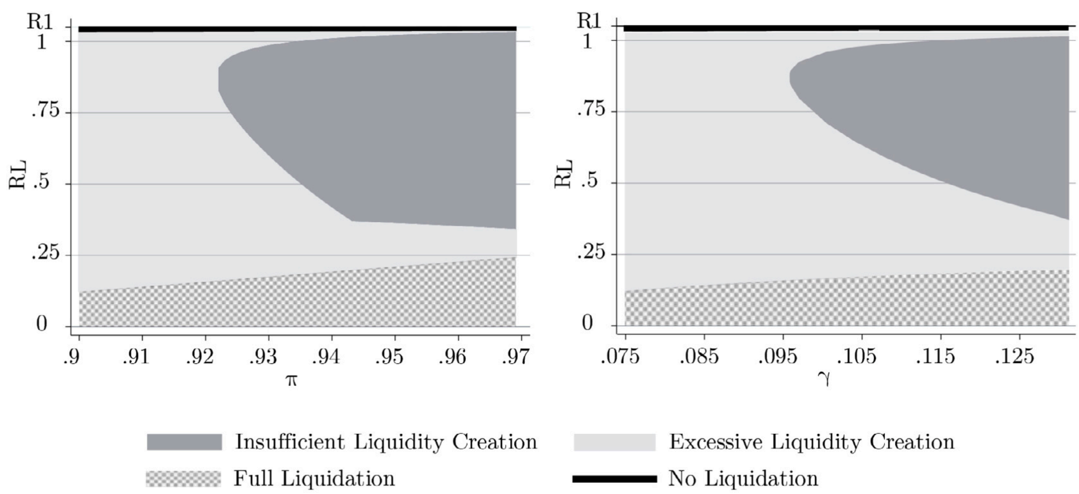

Fig. 3. Liquidity creation by π and γ. This figure shows the areas for which the wedge in liquidity creation $D(\mu^*) - D(\mu^{**})$ is positive (excessive liquidity creation), negative (insufficient liquidity creation) or zero (either no asset sales or full liquidation of all assets in both constrained efficient and equilibrium) for comparative statics with respect to the probability of receiving bad news π (left panel) and the liquidity benefit γ (right panel), both times in combination with R_L. Parametrization: $g(k) = \alpha ln(k)$ with $\alpha = 1.21$, $\gamma = 0.15$ (left panel), $\pi = 0.95$ (right panel), $R_H = 1.2$, $W = 1.2$ and $R_1 = 1.05$.

probability of "good news", π, i.e, when in economies in which fire sales are unlikely.

First, note that it is both socially optimal and privately optimal to rely more on market-based liquidity creation if the probability of a fire sale and thus the expected costs of the fire sale is lower. I.e., the intensity of fire sales—equivalent to the amount of assets sold—and thus the scope for insufficient liquidity creation increases in π. Said differently: for there to be insufficient liquidity creation, there need to be fire sales that involve large amounts of asset sales. For high π it is optimal for both socially and privately optimal to rely on market-based liquidity creation and thus more assets are sold in a fire sale. Hence, a fire sale will involve more asset sales if it was less likely to begin with, i.e., when π is high. The intensity of fire sales and thus the scope for insufficient liquidity creation is increasing in π.

3.5.2. Comparative statics with respect to γ

Next consider comparative statics with respect to γ. Recall that the parameter γ captures the liquidity benefit that is obtained from holding a safe asset. A higher γ implies that liquidity is valued more. The right panel of Fig. 3 plots as in the left panel the regions in which liquidity creation is excessive or insufficient. The figure illustrates that insufficient liquidity creation is more common in economies in which liquidity is valued relatively highly.

As with the comparative static with respect to π, we find that it is both privately and socially optimal to rely more on market-based liquidity creation as γ increases. Observe that the expected cost of a fire sale are independent of γ, but the benefits of liquidity creation are increasing in γ. Thus, the incentive to rely on market-based liquidity creation increases in γ for both social planner and individual intermediaries. However, as γ increases and as market-based liquidity creation becomes more attractive, this effect is stronger in the laissez-faire because in contrast to the social planner, competitive intermediaries do not internalize the pecuniary externalities of their asset sales that reduce welfare. Hence, as with π, the wedge between the constrained efficient and the laissez faire increases in γ, and thus the scope for insufficient liquidity creation increases as the valuation of liquidity γ increases.

3.5.3. Liquidity of the secondary market

As a final comparative static, we analyze the relation of secondary market liquidity on liquidity creation. As discussed above, the shape of late investors' production function $g(k)$ determines how fast the fire-sale price declines in the quantity of assets sold. In the following, we study two benchmark cases. First, we consider "cash-in-the-market" pricing (see, e.g., Allen and Gale, 1994; Parlatore, 2016) in which the asset price is determined by a fixed amount of liquidity available in the secondary market. Second, we consider a competitive asset market in which there is no price impact of asset sales. Note that both of these cases are nested in our model as special, limiting cases. Under cash-in-the-market pricing, the amount of liquidity created is always insufficient (similar to the findings in, e.g., Gale and Gottardi, 2015). In contrast, in the case of a competitive asset market, the wedge between the constrained-efficient and the equilibrium disappears. Both cases illustrate that insufficient liquidity creation is more likely to occur if secondary markets are not very deep and liquid.

Cash-in-the-market Pricing Under cash-in-the-market pricing, if a single intermediary increases the amount of assets sold, it decreases the fire-sale price while leaving the aggregate fire-sale revenue unchanged. I.e., it redistributes fire-sale revenues between intermediaries but does not increase the total liquidity generated by asset sales. However, increasing asset sales reduces hold-to-maturity liquidity creation. Thus, the social planner always prefer to stop selling assets as soon as a cash-in-the-market-pricing regime starts. At this point, any additional unit of asset sold reduces the liquidity. However, in the laissez-faire equilibrium, intermediaries must be indifferent between selling another unit or creating liquidity on balance sheet. This implies that the extent of the fire sale must induce a cash-in-the-market price that makes intermediaries indifferent between the two types of liquidity creation. While the overall amounts of funds transferred from investors to intermediaries are the same, there are more assets sold in the laissez faire and the fire-sale price is lower. This is why liquidity creation in the laissez faire is always insufficiently low under cash-in-the-market pricing.[10]

Competitive asset market In contrast, consider a competitive asset market in which the price is independent of quantities and the fire-sale discount is a constant. This is the case if and only if late investors' marginal productivity is constant, $\delta = 1/g'$. In such a competitive asset market, the wedge between the constrained-efficient and the equilibrium disappears. To begin with, the wedge occurs because the social planner internalizes the price-effect of asset sales whereas private agents do not–but if there is no price impact, there is nothing to internalize and the laissez-faire is already constrained efficient. Thus, there is no inefficiency in liquidation in a competitive market, and liquidity creation is not distorted either.[11]

[10] Our model nests a variant with cash-in-the-market pricing where investors always spend a constant amount $\tilde{M}$ on asset purchases: Consider a continuous and piecewise-linear function $g(k)$ with a kink at an input level $\tilde{k} = W - \tilde{M}$, and slope b for $k < W - \tilde{M}$ and slope $a < b$ for $k > W - \tilde{M}$. Recall that $\tilde{\delta}^{-1}$ describes the marginal productivity of the outside investors in the laissez-faire, so $\tilde{\delta}$ is the fire-sale discount at which banks are indifferent between held to maturity and market based liquidity creation. Let us assume that $\tilde{\delta}^{-1} \in [a, b]$, implying that g is sufficiently concave, i.e., the productivity is sufficiently high before and sufficiently low after the kink at $\tilde{k}$. Then the laissez-faire is characterized by $\delta^* = \tilde{\delta}$, $M^* = \tilde{M}$ and $\mu^* = \tilde{\delta}^{-1}\tilde{M}/R_1$, and the constrained efficient is characterized by $\delta^{**} = 1/b$, $M^{**} = \tilde{M}$ and $\mu^{**} = b\tilde{M}/R_1$. Thus, there is *de-facto* cash-in-the-market pricing: in both the constrained-efficient and in the laissez-faire, the outside investors invest $\tilde{k} = W - \tilde{M}$ in their production technology and use $\tilde{M}$ units for asset sales. Hence, while the overall amounts of funds transferred from investors to intermediaries is the same but more assets are sold in the laissez faire. Thus the amount of overall liquidity created is necessarily insufficient.

[11] Our model nests a competitive asset market in the form of a linear production function $g(k)$. Alternatively, we also converge to a competitive asset market as $W \to \infty$ as this makes the marginal productivity at the constrained-efficient de-facto constant. Because $g''(k) = 0$, the wedge between the constrained-efficient allocation and the laissez-faire disappears. The "pecuniary externality" term in (10) is zero in this case: $g'(W - M^{**}) - g'(W - M^*) = \frac{\gamma R_L M^{**} g''(W - M^{**})}{(1-\pi)R_1 + \gamma R_L} = 0$.

4. Policy implications

In this section, we discuss the model's implications for optimal regulation of financial intermediation. The scope of optimal regulation in our model is inherently macroprudential as excessive fire sales represent a systemic risk. I.e., the primary objective of optimal macroprudential regulation is to restrict the *contribution* of each intermediary to this systemic risk, rather than the *exposure* (as, e.g., in bank stress testing).

We consider three different regulatory approaches that aim at reducing the contribution of individual intermediaries to systemic risk: First, we consider standard tools of bank regulation that are based on banks' balance sheet measures, like equity and liquidity regulation. We refer to these as *leverage regulation*, and we will show that such policies fail if liquidity creation is too low. Second, we show that regulatory tools that target asset sales through quantity regulation or taxation directly could be effective, but such policy are not part of standard tool kit of regulators. Finally, we consider regulatory policies that are based on identify an intermediary's business model ex-ante and thus also work when leverage regulation fails. Moreover, by paying subsidies, this type of regulation can help to overcome problems of regulatory arbitrage that cannot be addressed by the first two tools.

Leverage Regulation We first analyze standard tools of banking regulation which are typically based on observable balance-sheet characteristics, and not on business models . For instance, capital requirements prescribe a minimum amount of risk-absorbing claims (equity, junior debt) as a function of the asset risk, thus targeting the solvency of banks. Likewise, liquidity requirements limit a bank's short-term liabilities as a function of its liquid assets, or require holding reserves as a fraction of deposits. Because assets are fixed, and because there are only two types of bank liabilities in our model (i.e., all debt is safe, "liquid" and short-term), the two types of regulations are equivalent in our model. Hence, we do not distinguish between solvency and liquidity regulation and refer to them simply as *leverage regulation*.

Crucially, restrictions on asset sales are not necessarily equivalent to leverage regulation in our model because the mapping between asset sales and the level of safe debt is not one-to-one.[12] This mechanic is illustrated in Fig. 4 which provides the same numerical example as Fig. 2 and again plots the level of safe debt D against the reliance on market-based intermediation μ. In this graph, the red horizontal lines illustrates how leverage regulation restricts the choices of banks, whereas the vertical green line represents asset-sales restrictions. If the aggregate safe-debt function $D(\mu)$ is non-monotone, there are aggregate levels of safe debt that correspond to two different levels of aggregate asset sales.

To illustrate the effectiveness or ineffectiveness of leverage regulation, we need to distinguish whether safe-debt creation is excessive or insufficient. If it is excessive (i.e., if $D(\mu^*) > D(\mu^{**})$), then we could use a macroprudential reserve requirement in order to implement μ^{**}, as suggested by Stein (2012): By choosing the required-reserve ratio and the amount of reserves in the system, the regulator can limit the level of safe debt created by all banks to $D(\mu^{**})$. The constrained-efficient level of market-based intermediation μ^{**} is the unique equilibrium outcome given this policy. The left panel of Fig. 4 is an illustration of this case: It holds that $D(\mu^*) > D(\mu^{**})$. Moreover, $D'(\mu^*) > 0$, i.e., there still is positive marginal liquidity creation in the laissez-faire. Note that this regulation would still be effective if there was "marginal liquidity destruction" in the competitive equilibrium (i.e., $D'(\mu^*) < 0$), as long as liquidity creation is excessive relative to the constrained-efficient

[12] Note that this mapping is one-to-one in the model by Stein (2012) because the lowest payoff of the bank's asset is zero. This implies that the liquidity regulation/reserve requirement proposed in Stein (2012) is actually equivalent to a regulation of asset sales in his model—but not in ours.

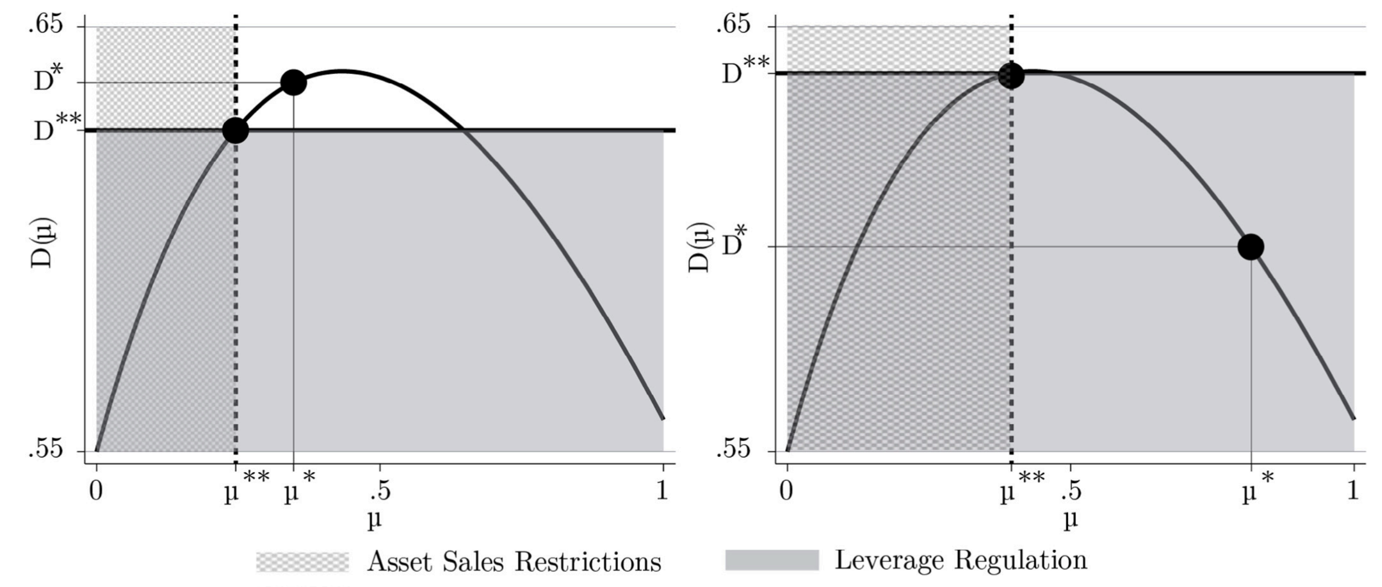

Fig. 4. Two scenarios for the aggregate liquidity creation in the constrained-efficient allocation and in the laissez-faire equilibrium. The solid black line represents the concave liquidity creation function $D(\mu)$ and is identical in both scenarios. Parametrization: $g(k) = \alpha ln(k)$ with $\alpha = 1.21$, $\gamma = 0.15$, $R_H = 1.2$, $W = 1.2$ and $R_1 = 1.05$. In the left panel ($\pi = 0.85$), there is excessive liquidity creation, whereas in the right panel ($\pi = 0.98$), there is not only liquidity destruction in equilibrium, but even insufficient liquidity creation compared with the constrained-efficient. The vertical dotted lines and the shaded areas to their left represent a regulation that limits asset sales. The horizontal solid lines and the gray area below them represent maximum leverage regulation. We see that while the two regulations are equivalent in the left case, the leverage regulation has no bite in the right case.

allocation. Both types of regulation are effective at implementing the constrained-efficient liquidation μ^{**} if liquidity creation excessive in the laissez-faire equilibrium.

The situation is different when the equilibrium level of safe debt in the unregulated economy is insufficiently low (i.e. if $D(\mu^*) \leq D(\mu^{**})$), as in the right panel of Fig. 4 provides the same numerical example as Fig. 2. Here, a macroprudential reserve requirement is ineffective as it is not binding. Note that a macroprudential leverage cap d^R can only binding if $d^R < D(\mu^*)$. However, such a cap cannot implement the constrained-efficient level $D(\mu^{**})$ because the laissez-faire level $D(\mu^*)$ is already inefficiently low, and a cap could then only reduce it further whereas the regulator would want to increase liquidity.

Hence, leverage regulation only allows to implement the constrained-efficient allocation if liquidity creation is excessive in the unregulated economy, but not if it is too low. In the latter case, the constrained-efficient allocation cannot be attained with standard tools of banking regulation, but only with direct regulatory tools like entry restrictions or taxation.

Finally, a subsidy on equity (or other loss-absorbing claims) works in the same way as leverage regulation, and is thus only effective if there is excessive liquidity creation in the laissez-faire equilibrium: By making equity funding cheaper, it becomes more attractive to adopt the hold-to-maturity business model which features a lower leverage. To subsidize equity, one could use tools like ACE (allowance on corporate equity) that aim to offset the tax shield of corporate taxation, where by choosing very high allowance rates, the tax shield can actually change its sign and become an equity subsidy.

The potential ineffectiveness of leverage regulation in our model should not be interpreted as an argument against the use of such regulatory tools in general. Our analysis on focuses on a specifically *macroprudential* objective. In reality, however, most regulatory tools (capital and liquidity regulation) have been introduced primarily with a *microprudential* objective. As our model does not feature moral hazard (nor any other agency friction), we only have a macroprudential objective and thus abstract from potential microprudential problems for which leverage regulation may be useful.

Quantity Regulation and Pigouvian Taxation

Having established that the standard tools of bank regulation are not effective in case of insufficient liquidity creation, we now turn towards two direct regulatory tools through which the constrained-efficient allocation can be implemented: Either the aggregate amount of assets sold can be limited to μ^{**}—a quantity regulation—or a transaction tax on assets sales can be imposed—a "Pigouvian taxation" or "prudential taxation" approach. Such policies are substantially different from the standard tools of banking regulation as they are not based on observable balance-sheet characteristics, but on the banks business models directly.

Using quantity regulation, a regulator can implement the constrained-efficient allocation by giving out permits to engage in market-based intermediation to a number of μ^{**} intermediaries. If there was unobservable heterogeneity across intermediaries, the regulator could also use a cap-and-trade approach: Intermediaries initially receive permits specifying how much reliance on market-based banking is allowed (i.e., how much assets may be sold), with the sum of these permits being μ^{**}. As intermediaries are allowed to trade these permits, market-based intermediaries would purchase permits from hold-to-maturity intermediaries. Effectively, such a regulation would restrict the entry of intermediaries into the market of market-based intermediation.

If the regulator prefers to use a price regulation instead, the corresponding approach would be a transaction tax on asset sales ("Pigouvian tax"), the proceeds of which would be redistributed to intermediaries in a lump-sum fashion. Intermediaries and consumers would recognize how this changes their ability to create liquidity and to make profits, and intermediation would shift towards the hold-to-maturity sector. A formal analysis of this taxation approach can be found in Appendix E.

Regulatory Arbitrage An important caveat of the types of regulatory tools discussed above is the threat of regulatory arbitrage (Buchak et al., 2018). Regulatory requirements create an incentive to shift intermediation activities outside of the regulatory perimeter if possible. Intermediaries could hence attempt to circumvent the regulatory requirements by conducting market-based intermediation in entities off their balance sheet and outside of the regulatory perimeter. The history of financial regulation is full of examples how the financial industry adjusts and circumvents regulatory requirements, such as the creating money market mutual funds to circumvent deposit-rate regulation (Regulation Q) or the granting of "liquidity guarantees" to off-balance-sheet subsidiaries prior to the global financial crisis is another recent example (Acharya et al., 2013). In the following, we will therefore analyze tools of macroprudential regulation that are immune to regulatory arbitrage.

Assuming that regulatory arbitrage is possible, i.e., assuming that it is possible to conduct market-based liquidity creation outside of the regulatory perimeter in a "shadow banking" sector, it becomes apparent that direct restrictions on market-based financial intermediation are rendered ineffective. In particular, if arbitrage is possible at little or not cost, any regulatory measure targeting market-based intermediation will be avoided by migrating the intermediation activity to the shadow banking sector.

However, while dis-incentivizing market-based liquidity may be ineffective in the light of regulatory arbitrage, incentivizing hold-to-maturity banking may allow the social planner to nonetheless implement the constrained-efficient allocation (Ordoñez, 2018). In particular, we argue that a Pigouvian subsidy for hold-to-maturity banking is isomorphic to a Pigouvian tax on market-based intermediation, but in contrast to the latter, the former is effective even if regulatory arbitrage is possible.

The problem is that under the constrained-efficient sector sizes, asset sales are low and the fire-sale price is high, so profits are larger in the market-based sector, and intermediaries have an incentive to enter this sector. If the social planner has the ability to raise taxes (for example by taxing consumption), it could use these funds to pay a subsidy to hold-to-maturity intermediaries. To make this work, it would have to introduce a regulation that specifies that an intermediary is only eligible for this subsidy if its business model does not involve any asset liquidation. Supervision might also be necessary in order to enforce the regulation. The required subsidy is $S^{**} = \Pi_{MB}(\delta^{**}) - \Pi_{HTM}$, it has to be equal to the difference between the profit of market-based intermediation profit given the constrained-optimal aggregate liquidation level, and the hold-to-maturity intermediation profit (which is independent of the fire-sale discount factor δ and thus of μ), compare (6) and (7).

A tax on market-based banking which cannot be applied in a shadow-banking sector has the effect of pushing intermediation outside of the regulatory perimeter, i.e., it encourages regulatory arbitrage. A subsidy, in contrast, increases the attractiveness of complying with regulatory standards. The downside of a subsidy is of course that it is a cost for the government.

How could such a subsidy be implemented in practice? In addition to a one-to-one implementation of a direct subsidy for a specific business model, we want to briefly discuss additional implementation methods. Such implementation methods could take the form of explicit or implicit guarantees, or of a subsidy for equity or risk-absorbing long-term funding. Any kind of implicit or explicit guarantee that commercial banks receive from the government or central banks, like implicit bailout guarantees or the access to a discount window, can also have the effect of subsidizing the hold-to-maturity relative to market-based banking performed by shadow banks. Another example of such a subsidy is an underpriced deposit insurance. The idea to use an underpriced deposit insurance as a policy tool to improve the efficiency of banking has been made in other contexts, see for example Morrison and White (2011). The implementation, however, is not straightforward in the context of the stylized model that we have presented because in this model there is no role for a deposit insurance at all. A deposit insurance is only valuable if there is a risk of runs on financial intermediaries, or, and this is potentially more relevant, if there is some "disaster risk" that the deposit insurance can protect a hold-to-maturity intermediary against. In Appendix F, we briefly show that if there is some very small probability in which the asset payoff is even below R_L, a deposit insurance that would step in this specific case would allow to maintain a safe-debt level of R_L. Such a deposit insurance could be offered below the actuarially fair price to hold-to-maturity institutions (maybe even free of charge) to implement the proposed subsidy.

Alternatively, one could also set up a deposit insurance just for hold-to-maturity intermediaries that guarantees a repayment slightly above R_L. This deposit insurance would then pay a small amount to depositors whenever state L realizes and thus increases the "safe-debt capacity" of hold-to-maturity intermediaries and make this business model attractive relative to market-based intermediation.

5. Discussion

Recent decades have seen the rise of new, market-based forms of financial intermediation. While the empirical understanding has made progress, there is little normative guidance from economic theory on the optimality of these market-based forms of intermediation. This paper attempts to fill this gap by providing a simple theory on the coexistence of traditional and market-based financial intermediation. This coexistence theory is complementary to the explanations of shadow banking as regulatory arbitrage (see, e.g., Buchak et al., 2018). In the theory proposed, traditional banks create liquidity by issuing equity and holding assets to maturity, and are never subject to withdrawals. In contrast, market-based intermediaries create liquidity by selling assets in fire sales and can be subject to withdrawals in downturns.

Our model has two key implications. First, there is a natural tendency for the extent of market-based forms of intermediation to be excessive. Importantly, market-based intermediation is excessive even though it is not a response to regulatory requirements, as typically assumed in the literature on shadow banking. Therefore, our model suggests that the natural degree of market-based intermediation and the resulting fire sales may be excessive even when abstracting from the possibility of regulatory arbitrage.

Second, the excessive reliance on market-based banking implies that liquidity creation is generally inefficient. While the degree of market-based intermediation is always excessive, the amount of safe debt created can be excessive as well, but it can also be insufficiently low. While individual intermediaries may have incentives to create safe debt by relying on fire sale in downturns, in aggregate, the excessive reliance on market-based intermediation can induce a fire-sale discount that pushes the amount of safe debt below its optimal level. I.e., our model suggests that it is possible for the financial sector to excessively rely on liquidity creation via shadow banking or by market-based activities from financial institutions. In turn, the arising fire sale risk may reduce the amount of safe assets generated below its optimal level.

Thereby, our model has important implications for the macroprudential regulation of financial intermediation. We point out that standard instruments of banking regulation, such as capital and liquidity regulation, are not necessarily effective in closing the wedge between the constrained-efficient allocation and the equilibrium. This contrasts with a number of recent theoretical contributions that propose to address pecuniary externalities using liquidity regulation (Stein, 2012), capital requirements (Gersbach and Rochet, 2012, 2017; Malherbe, 2020), or a mix of both (Walther, 2016).

In contrast, our model suggests that optimal macroprudential regulation should target the composition of the financial intermediation system and hence distinguish between specific business models of financial intermediation rather than focusing on balance sheet measures. Given that market-based intermediation has a natural tendency to be excessive, we argue that the most effective policy is to subsidize the traditional, hold-to-maturity bank business model. Importantly, this policy allows to effectively control the aggregate composition of financial intermediation and even address the threat of regulatory arbitrage.

CRediT authorship contribution statement

Stephan Luck: Conceptualization, Methodology, Software, Formal analysis, Writing. **Paul Schempp:** Conceptualization, Methodology, Formal analysis, Writing, Visualization.

Data availability

No data was used for the research described in the article.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jfi.2022.100996.

References

Acharya, Viral V., Schnabl, Philipp, Suarez, Gustavo A., 2013. Securitization without risk transfer. J. Financ. Econ. 107 (3), 515–536.

Admati, Anat R., Hellwig, Martin F., 2013. The Bankers' New Clothes: What's Wrong with Banking and What To Do About It. Princeton University Press, Princeton, NJ.

Allen, Franklin, Gale, Douglas, 1994. Limited market participation and volatility of asset prices. Amer. Econ. Rev. 84 (4), 933–955.

Ari, Anil, Darracq-Paries, Matthieu, Kok, Christoffer, Dochowski, Dawid, 2017. Shadow banking and market discipline on traditional banks. IMF Work. Pap. 17 (285), 1.

Bianchi, Javier, Mendoza, Enrique G., 2018. Optimal time-consistent macroprudential policy. J. Polit. Econ. 126 (2), 588–634.

Buchak, Greg, Matvos, Gregor, Piskorski, Tomasz, Seru, Amit, 2018. Fintech, regulatory arbitrage, and the rise of shadow banks. J. Financ. Econ. 130 (3), 453–483.

Cetorelli, Nicola, Mandel, Benjamin H., Mollineaux, Lindsay, 2012. The evolution of banks and financial intermediation: Framing the analysis. Econ. Policy Rev. 18 (2), 1–12.

Chrétien, Edouard, Lyonnet, Victor, 2019. Traditional and Shadow Banks. SSRN Working Paper 3376891.

Dang, Tri Vi, Gorton, Gary, Holmström, Bengt, 2015. Ignorance, Debt and Financial Crises. Working Paper.

Dang, Tri Vi, Gorton, Gary B., Holmström, Bengt, Ordoñez, Guillermo, 2017. Banks as secret keepers. Amer. Econ. Rev. 107 (4), 1005–1029.

Dávila, Eduardo, Korinek, Anton, 2018. Pecuniary externalities in economies with financial frictions. Rev. Econom. Stud. 85 (1), 352–395.

Diamond, Douglas W., Dybvig, Philip H., 1983. Bank runs, deposit insurance, and liquidity. J. Polit. Econ. 91 (3), 401–419.

Financial Crisis Inquiry Commission, 2011. The Financial Crisis Inquiry Report. Printing Office of the US Government, Washington D.C..

Gale, Douglas, Gottardi, Piero, 2015. Capital structure, investment, and fire sales. Rev. Financ. Stud. 28 (9), 2502–2533.

Gersbach, Hans, Rochet, Jean-Charles, 2012. Aggregate investment externalities and macroprudential regulation. J. Money Credit Bank. 44 (Supplement to 2), 73–109.

Gersbach, Hans, Rochet, Jean-Charles, 2017. Capital regulation and credit fluctuations. J. Monetary Econ. 90, 113–124.

Gorton, Gary B., 2010. Slapped By the Invisible Hand: The Panic of 2007. Oxford University Press.

Gorton, Gary B., Metrick, Andrew, 2010. Regulating the shadow banking system. Brook. Pap. Econ. Act. 261–312.

Gorton, Gary B., Metrick, Andrew, 2012. Getting up to speed on the financial crisis: a one-weekend-reader's guide. J. Econ. Lit. 50 (1), 158–178.

Gorton, Gary B., Pennacchi, George, 1990. Financial intermediaries and liquidity creation. J. Finance XLV (1), 49–71.

Hanson, Samuel G., Shleifer, Andrei, Stein, Jeremy C., Vishny, Robert W., 2015. Banks as patient fixed-income investors. J. Financ. Econ. 117 (3), 449–469.

Hellwig, Martin F., 2015. Existence and Constrained (In-)Efficiency of General Competitive Equilibrium in a Model of Liquidity Provision and Equity Funding of Banks under Uncertainty. Mimeo, Bonn.

Holmström, Bengt, Tirole, Jean, 1998. Private and public supply of liquidity. J. Polit. Econ. 106 (1), 1–40.

Holmström, Bengt, Tirole, Jean, 2011. Inside and Outside Liquidity. The MIT Press, Cambridge, MA.

Huang, Ji, 2018. Banking and shadow banking. J. Econom. Theory 178, 124–152.

Kara, Gazi I., Ozsoy, S. Mehmet, 2020. Bank regulation under fire sale externalities. Rev. Financ. Stud. 33 (6), 2554–2584.

Krishnamurthy, Arvind, Nagel, Stefan, Orlov, Dmitry, 2014. Sizing up repo. J. Finance LXIX (6), 2381–2417.

Lorenzoni, Guido, 2008. Inefficient credit booms. Rev. Econom. Stud. 75 (3), 809–833.

Malherbe, Frederic, 2020. Optimal capital requirements over the business and financial cycles. Am. Econ. J. Macroecon. 12 (3), 139–74.

Martinez-Miera, David, Repullo, Rafael, 2019. Markets, Banks, and Shadow Banks. ECB Working Paper No 2234.

Morrison, Alan D., White, Lucy, 2011. Deposit insurance and subsidized recapitalizations. J. Bank. Financ. 35 (12), 3400–3416.

Ordoñez, Guillermo, 2018. Sustainable shadow banking. Am. Econ. J. Macroecon. 10 (1), 33–56.

Parlatore, Cecilia, 2016. Fragility in money market funds: Sponsor support and regulation. J. Financ. Econ. 121 (3), 595–623.

Plantin, Guillaume, 2015. Shadow banking and bank capital regulation. Rev. Financ. Stud. 28 (1), 146–175.

Pozsar, Zoltan, Adrian, Tobias, Ashcraft, Adam, Boesky, Hayley, 2013. Shadow banking. In: Federal Reserve Bank of New York Economic Policy Review. (458), pp. 1–16.

Segura, Anatoli, 2018. Why did sponsor banks rescue their sivs? A signaling model of rescues. Rev. Finance 22 (2), 661–697.

Stein, Jeremy C., 2012. Monetary policy as financial stability regulation. Q. J. Econ. 127 (1), 57–95.

Walther, Ansgar, 2016. Jointly optimal regulation of bank capital and liquidity. J. Money Credit Bank. 48 (2–3), 415–448.