American Economic Review 2019, 109(1): 271–313
https://doi.org/10.1257/aer.20161488

Optimal Regulation of Financial Intermediaries[†]

By Sebastian Di Tella*

I characterize the optimal financial regulation policy in an economy where financial intermediaries trade capital assets on behalf of households, but must retain an equity stake to align incentives. Financial regulation is necessary because intermediaries cannot be excluded from privately trading in capital markets. They don't internalize that high asset prices force everyone to bear more risk. The socially optimal allocation can be implemented with a tax on asset holdings. I derive a sufficient statistic for the externality and use market data on leverage and volatility of intermediaries' equity to measure it. (*JEL* D82, G01, G12, G20, G31, H25)

Financial intermediaries play an important role in modern economies, trading capital assets on behalf of households. However, excessive risk taking by financial intermediaries can create macro instability and lead to financial crises. This has created a large interest in the regulation of financial intermediaries, especially after the 2008 financial crisis. But what are the right policy instruments, and what is the optimal way to use them?

In this paper I study the optimal financial regulation policy in a macroeconomic model of financial crises where intermediaries trade capital on behalf of households, but must retain an equity stake to align incentives. This is a commonly observed financial arrangement, and widely used in the literature.[1] Hidden trade is an important feature of the environment. The equity constraint comes from a moral hazard problem with hidden trade: intermediaries can divert investment returns. Because their job is precisely to buy and sell capital assets, they cannot be excluded from privately trading in this market. This is the ultimate source of inefficiency in this economy.[2] In contrast to previous papers, I don't put any constraints on private contracts. I compare the competitive equilibrium where agents can write complete *long-term* contracts with the best allocation that can be achieved by a social planner facing the same informational frictions. An advantage

* Stanford GSB, 655 Knight Way, Stanford, CA 94305 (email: sditella@stanford.edu). This paper was accepted to the *AER* under the guidance of John Leahy, Coeditor. I'd like to thank Andres Schneider, Andy Skrzypacz, Peter DeMarzo, Pablo Kurlat, Yuliy Sannikov, Bob Hall, Martin Schneider, Monika Piazzesi, V.V. Chari, Chad Jones, Chris Tonetti, Florian Scheuer, Eric Madsen, Alex Bloedel, Javier Bianchi, Fernando Alvarez, Takuo Sugaya, Victoria Vanasco, Mike Harrison, and Peter Kondor.

[†] Go to https://doi.org/10.1257/aer.20161488 to visit the article page for additional materials and author disclosure statement(s).

[1] See Brunnermeier and Sannikov (2014), He and Krishnamurthy (2012), and Di Tella (2017).

[2] It is well known that hidden trade can be a source of inefficiency (see, for example, Farhi, Golosov, and Tsyvinski 2009 and Kehoe and Levine 1993). The contribution in this paper is to characterize the optimal financial regulation policy in a widely used and policy-relevant model of financial crises.

of this mechanism-design approach is that I don't need to commit to an arbitrary set of policy instruments: I let the model guide the choice of policy instrument.

The main takeaway is that the socially optimal allocation requires lower asset prices in order to reduce intermediaries' exposure to risk, even if it comes at the cost of lower investment. Essentially, lower asset prices make the whole financial system less risky. I show that the socially optimal allocation can be implemented as a competitive equilibrium with a tax on asset holdings that reduces asset prices. Once we do this, there is no need for further regulation that distorts intermediaries' decisions (e.g., capital requirements). In particular, the unregulated competitive equilibrium may feature a financial amplification channel with aggregate risk excessively concentrated on the balance sheets of intermediaries, compared to the socially optimal allocation. But in contrast to most of the literature, this is not the result of incomplete markets. Intermediaries optimally choose this risk exposure, and privately optimal contracts agree with the planner on the cost of delivering utility to intermediaries across aggregate states, conditional on the rest of the allocation. This is a feature of long-term contracts that does not hold with short-term contracts. As a result, there is no need to directly regulate intermediaries' exposure to aggregate risk (e.g., bailouts, stress tests).

To understand why an intervention is necessary, notice that hidden trade creates an externality because the private benefit of diverting investment returns depends on the market value of capital assets. Intermediaries don't internalize that by demanding capital and bidding up its price, they worsen the moral hazard problem for everyone else. As a result, asset prices are too high in equilibrium, and intermediaries must take too much idiosyncratic risk in order to align incentives. The optimal policy targets the source of the inefficiency directly (high asset prices) without distorting other margins.

The hidden-trade externality admits a sufficient statistic representation in terms of measurable variables, valid for heterogeneous intermediaries and asset classes, and arbitrary aggregate shocks. The unregulated competitive equilibrium has Marginal cost of capital $=$ Marginal value of capital, as usual. However, because of the externality produced by hidden trade, the planner's first-order condition (FOC) for investment is actually Marginal cost of capital $\times (1 + \eta_t) =$ Marginal value of capital. Higher investment requires higher asset prices, which worsen the moral hazard problem and increase intermediaries' exposure to risk. The externality $\eta_t \geq 0$ measures this additional marginal cost of capital, and can be written

$$\eta_t = \alpha_t \epsilon_t, \tag{1}$$

where α_t is intermediaries' equilibrium risk-adjusted expected excess return on assets, and ϵ_t the technologically given elasticity of the cost of capital. It has a simple interpretation. If we want to raise investment, asset prices (equal to the marginal cost of capital) will have to increase by ϵ_t percent. As a result, intermediaries will have to increase the value of their asset holdings, exposing them to more idiosyncratic risk. In equilibrium, the excess return α_t compensates intermediaries for the idiosyncratic risk they must take when holding capital, so $\eta_t = \alpha_t \epsilon_t$ measures the additional marginal cost of capital coming from intermediaries' larger exposure to idiosyncratic risk. We can implement the optimal allocation by setting the present

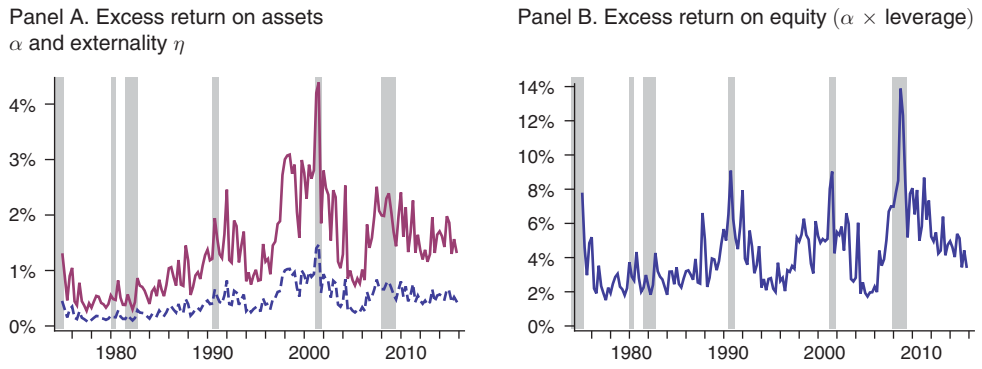

FIGURE 1

Notes: Panel A: intermediaries' risk-adjusted expected excess return on assets α (dashed) and externality η (solid). Panel B: intermediaries' risk-adjusted expected excess return on equity $\alpha \times$ leverage.

value of the tax on asset holdings (relative to the market value of assets) equal to the externality, $T_t/q_t = \eta_t$.[3] Intuitively, this reduces the market price of capital, internalizing the externality.

Expression (1) is valid for any type of aggregate shock, and allows us to measure the externality using market data without having to specify many structural features of the model. The model also provides a simple endogenous formula for the excess return α_t in terms of intermediaries' leverage and idiosyncratic risk, which may be easier to measure and allows us to understand how the externality is affected by aggregate shocks. As we would expect, the equilibrium excess return α_t is large when intermediaries' idiosyncratic risk and leverage are high, or when financial constraints are tight, e.g., during financial crises.

I build a time series for the model-predicted excess return α_t and the externality η_t using CRSP/Compustat data. Figure 1 shows the resulting time series for the baseline calibration. The average externality is 1.36 percent, and it can go up to 4.39 percent during downturns when the excess return on assets α_t is large. To put this in context, a 1.36 percent reduction in the marginal cost of capital corresponds to a reduction in the investment share of GDP of roughly 1.36 percentage points (e.g., from 20 percent to 18.64 percent of GDP). The model also yields an average excess return on equity (leverage$_t \times \alpha_t$) for financial intermediaries of 4.26 percent, which also spikes during downturns: up to 13.88 percent at the peak of the financial crisis.[4]

An important practical question for regulators is how different intermediaries and different asset classes should be treated (e.g., how should we risk-weight different assets?). To address this issue, I extend the baseline model to incorporate heterogeneous intermediaries and asset classes. The optimal allocation can still be

[3] The optimal policy will in general require a time-varying tax rate, but it can be implemented by continuously adjusting the tax rate to target the present value T_t.

[4] This is the gross alpha on equity, before fees. In the model outside investors don't obtain any excess return.

implemented with taxes on asset holdings. Each asset class has its own tax, but we do not discriminate between intermediaries. The sufficient statistic (1) is true for each asset class and tells us how each class should be treated. Regulators should not be directly concerned with the risk of each asset class, or even their systemic risk. Rather, for each asset class j, the average excess return $\alpha_{j,t}$ contains all the relevant information (along with the elasticity ϵ_j), and reflects both the asset class' risk and the place it occupies on intermediaries' balance sheets.

Finally, it is worth contrasting the tax on assets with capital requirements, which are a common component of financial regulation policy in practice. The tax makes assets less attractive, but doesn't affect the debt/equity margin. Intermediaries have private reasons for preferring a certain debt and equity mix, related to insiders' incentives, and the planner doesn't need to interfere. In contrast, capital requirements penalize the use of debt, so they unnecessarily interfere in this margin. To the extent that distorting the debt/equity margin is costly for intermediaries, capital requirements also act as a tax on assets, but only indirectly. As a result, while capital requirements could be used in a welfare improving way by a social planner, they are not the optimal policy instrument in this environment.[5]

I use a continuous-time setup with Epstein-Zin (EZ) preferences and arbitrary aggregate shocks that allows me to connect results with asset pricing literature. A methodological contribution of the paper consists of characterizing optimal dynamic contracts in this environment with recursive EZ preferences and arbitrary prices or exogenous shocks. The competitive equilibrium and the social planner's allocation can be characterized with a system of partial differential equations (PDEs). I describe the procedure in the online Appendix. I also provide discrete-time version of the setting in the online Appendix.

Literature Review.—I use an environment similar to Brunnermeier and Sannikov (2014), Di Tella (2017), and He and Krishnamurthy (2012, 2013), where financial intermediaries trade capital on behalf of households but must retain an equity stake for incentive reasons.[6] Whereas these papers' main contribution is a positive explanation of financial crises, the main contribution of this paper is the characterization of the optimal regulation policy in this environment. In order to understand the efficiency of the competitive equilibrium and the scope for financial regulation, it is important that we don't impose any ad hoc constraints on the contract space. In contrast to this literature, I allow private agents to write complete, long-term contracts with full commitment. While technically more involved, long-term contracts play an important role in the welfare and policy analysis. In contrast, most of these papers feature incomplete short-term contracts, where intermediaries are not allowed to share aggregate risk. A planner can therefore improve the competitive allocation by either completing the market, or by redistributing wealth through asset prices. However, financial markets provide ample opportunities for financial

[5] Capital requirements may be justified if government bailouts create incentives for banks to take too much risk (Chari and Kehoe 2013). In this environment there is no reason for bailouts, however.

[6] More generally, this paper fits into the literature on financial amplification channel going back to Kiyotaki and Moore (1997) and Bernanke, Gertler, and Gilchrist (1999).

intermediaries to hedge aggregate risk. Di Tella (2017) does allow contingent contracts and shows that the competitive equilibrium may nonetheless feature a financial amplification channel. The mechanism behind the financial amplification channel here is a generalized version of the mechanism in that paper. However, Di Tella (2017) still looks only at short-term contracts, and this turns out to be crucial for the purpose of financial regulation, which is the focus of this paper. Private short-term contracts do not internalize that giving intermediaries more wealth can relax the risk sharing problem, so a planner that can only regulate intermediaries' exposure to aggregate risk can improve the competitive allocation, even though private agents are free to share aggregate risk. Long-term contracts eliminate this source of inefficiency. In addition, as it turns out, optimal long-term contracts are renegotiation proof.

The ultimate source of inefficiency in this environment is that financial intermediaries' activity in capital markets cannot be easily monitored. It is well known that hidden trade has the potential to introduce inefficiency into a competitive equilibrium, as in Farhi, Golosov, and Tsyvinski (2009) or Kehoe and Levine (1993), because a social planner may be able to indirectly affect the equilibrium price in the hidden market and relax the incentive constraints.[7] The contribution of this paper is to characterize the resulting externality and optimal policy in a widely used and policy-relevant class of models of financial crises.

A different strand of the literature emphasizes incomplete markets as a source of inefficiency (see Hart 1975, Geanakoplos and Polemarchakis 1986, Stiglitz 1982, Geanakoplos et al. 1990). In Lorenzoni (2008) and Korinek (2012), more productive agents are endogenously unable to obtain enough insurance against downturns because of a limited commitment problem, so the marginal rates of substitution don't equalize. Raising the equilibrium price of assets that more productive agents hold is a way of transferring resources to them. He and Kondor (2016) and Dávila et al. (2012) also feature this source of inefficiency. Here I assume complete contracts/markets and the moral hazard problem does not restrict aggregate risk sharing, so this source of inefficiency is absent. In this line, Rampini and Viswanathan (2010) study an economy with borrowing constraints derived from a limited commitment problem, but allow complete financial markets. Credit-constrained firms may decide to forgo insurance against aggregate shocks in order to obtain more funds to invest up front. Since financial markets are complete, however, this aggregate risk sharing is efficient. Alvarez and Jermann (2000) derive borrowing constraints from a limited commitment problem. Since there is no restriction on contracts, and no hidden trade, the competitive equilibrium is efficient.

Inefficiency can also arise when financial frictions have prices in them, since private agents may not internalize how their actions affect those constraints through prices, such as in Bianchi (2011) and Bianchi and Mendoza (2011). Gersbach and Rochet (2012) study an economy where bankers face a moral hazard problem, and show that a planner who limits liquidations can improve ex ante

[7] It is worth clarifying the role of decentralized contracts: if all agents in the economy could get together and write one large contract, we would obtain efficiency; this is in fact the planner's problem. The competitive equilibrium is inefficient in the sense that the planner's allocation cannot be decentralized with a competitive market without any policy intervention.

social welfare. Farhi and Werning (2016) instead look at economies with nominal rigidities, where prices are fixed but Keynesian aggregate demand effects create a scope for regulation. In most of this literature the planner's policy instruments are restricted. In contrast, here I characterize the best allocation that can be achieved by a social planner facing the same informational frictions. This allows me to use the model to guide the choice of policy instrument. In general, when the inefficiency is caused by a price, ideally we would want to use an instrument that affects the price and as few other things as possible, to the extent such an instrument is available.

The contractual setting is related to the partial equilibrium settings in Sannikov (2008), DeMarzo and Sannikov (2006), He (2012), DeMarzo et al. (2012), and Biais et al. (2007). In particular, I use the same contractual setting as in Di Tella and Sannikov (2016) who characterize optimal contracts in a stationary environment where agents have access to hidden savings. A methodological contribution of this paper consists of characterizing optimal dynamic contracts in this environment with recursive EZ preferences and arbitrary prices or exogenous shocks. I rule out hidden savings, but the impact of hidden savings on the optimal financial regulation policy seems like a fruitful avenue for future research.

I. The Model

I build on the models of financial crises in Brunnermeier and Sannikov (2014), Di Tella (2017), and He and Krishnamurthy (2012, 2013). The main difference is that I allow agents to write fully contingent long-term contracts, and I consider arbitrary aggregate shocks that can affect any feature of the environment.

A. *Setting*

Technology.—The economy is populated by a continuum of households and financial intermediaries, identical in every respect except that intermediaries can trade capital on behalf of households. There are two goods, consumption and capital. Denote by k_t the aggregate "efficiency units" of capital in the economy, and by $k_{i,t}$ the individual holdings of intermediary i, where $t \in [0, \infty)$ is time. Capital can be costlessly reallocated between intermediaries, so $k_{i,t}$ will be a choice variable. However, capital is exposed to both aggregate and intermediary-specific idiosyncratic risk. If an intermediary holds $k_{i,t}$ units of capital, he gets a "capital quality" shock[8]

$$d\Delta_{i,t} = \sigma_t k_{i,t} dZ_t + \nu_t k_{i,t} dW_{i,t},$$

[8] In other words, while $k_{i,t}$ is a choice variable, the cumulative change in the capital stock for which intermediary i is responsible up to time t is $\int_0^t d\Delta_{i,t}$.

where Z is an aggregate d-dimensional Brownian motion, and W_i is an idiosyncratic Brownian motion for each intermediary i.[9] Here Z represents an aggregate total factor productivity (TFP) shock, and W_i the outcome of intermediary i's idiosyncratic activity.[10]

Capital produces a flow of consumption goods ak_t. In addition, competitive investment firms use capital to produce a flow of new capital $g_t k_t$ at a cost $\iota_t(g_t)\, k_t$ in consumption goods, where $\iota_t' \geq 0$ and $\iota_t'' \geq 0$. As a result of investment and shocks, the aggregate capital stock k follows the law of motion,

$$\frac{dk_t}{k_t} = g_t dt + \sigma_t dZ_t,$$

where the idiosyncratic shocks W_i have been aggregated away.

We can let several features of the environment, such as σ_t, ν_t, $\iota_t(\cdot)$, or ϕ_t introduced below, depend on the history of aggregate shocks Z. To this end introduce an exogenous aggregate state of the economy $Y_t \in \mathbb{R}^n$, with law of motion,

$$dY_t = \mu_Y(Y_t)dt + \sigma_Y(Y_t)\, dZ_t.$$

We can later specify how this aggregate state affects the economy, e.g., $\nu_t = \nu(Y_t)$ for uncertainty shocks, or $\iota_t(g) = \iota(g; Y_t)$ for shocks to investment technology. Notice that Y is driven by the same Z we called a TFP shock above, but this is without loss of generality because Z and Y can be multidimensional. TFP shocks to k may or may not be correlated with shocks to other features of the environment.

Preferences.—Both intermediaries and households have Epstein-Zin preferences with the same discount factor ρ, risk aversion γ, and elasticity of intertemporal substitution (EIS) ψ:

$$U_t = E_t\left[\int_t^\infty f(c_u, U_u)\, du\right], \tag{2}$$

where the EZ aggregator takes the form

$$f(c, U) = \frac{1}{1 - 1/\psi}\left\{\frac{c^{1-1/\psi}}{[(1-\gamma)U]^{\frac{\gamma-1/\psi}{1-\gamma}}} - \rho(1-\gamma)U\right\}.$$

[9] Z and $\{W_i\}_{i\in\mathbb{I}}$ are all mutually independent and admit an exact law of large numbers. See Sun (2006) for details.

[10] For example, if two intermediaries invest \$1 they will obtain different returns depending on their precise investment strategies.

I will focus on the case where relative risk aversion is larger than log: $\gamma > 1$, and elasticity of intertemporal substitution is larger than 2: $\psi > 2$.[11,12]

Markets and Investment.—There is a complete financial market with risk-free rate r and price π for aggregate risk Z. Idiosyncratic risks W_i are tradable but have zero price in the financial market since they can be aggregated away. Let Q be the equivalent martingale measure associated with r and π.

There is a competitive market for capital with price $q > 0$ with law of motion,

$$\frac{dq_t}{q_t} = \mu_{q,t}dt + \sigma_{q,t}dZ_t.$$

Investment firms rent capital from intermediaries to produce new capital. Their profit maximization yields Tobin's q,

$$\iota_t'(g_t) = q_t,$$

and a rental price for capital $r_t^k = q_t g_t - \iota_t(g_t)$.[13]

Prices q, r, π depend on the history of aggregate shocks Z and are determined in equilibrium.

Tax on Asset Holdings.—I will later show that the planner's optimal allocation can be implemented with a tax on assets, so it is useful to introduce it at this point. An intermediary who holds capital worth $q_t k_{i,t}$ must pay a tax flow $\tau_t^k q_t k_{i,t}$, where τ^k may depend on the history of aggregate shocks Z. As a result the government raises a flow $\tau_t^k q_t k_t$, which can be distributed back to agents via lump-sum transfers. The present value of transfers is $T_t k_t$,

$$T_t = \frac{1}{k_t} E_t^Q \left[\int_t^\infty e^{-\int_t^u r_m dm} \tau_u^k q_u k_u \, du \right]. \tag{3}$$

These transfers are part of private agents' aggregate wealth $(q_t + T_t)k_t$. In the unregulated economy, we simply take $\tau^k = 0$ and therefore $T = 0$. It should be stressed that I am not restricting the planner to this policy instrument, but rather finding that the optimal allocation can be implemented this way.

[11] It is natural to focus on the case with elasticity of intertemporal substitution greater than 1, especially when studying economies with stochastic volatility. The further restriction to EIS > 2 is required to ensure the existence of the competitive equilibrium. See the discussion below on hidden savings. The empirical literature on the EIS is mixed. Several authors find an EIS less than 1 (Hall 1988, Vissing-Jorgensen 2002) while others an EIS of 1.5, 2, or even larger (Beeler and Campbell 2009, Bansal et al. 2014, Gruber 2013, Mulligan 2002).

[12] For applications it might be useful to introduce retirement among intermediaries, which arrives with Poisson intensity θ, in order to obtain a stationary distribution. I allow for this in the online Appendix. For simplicity we can focus on $\theta = 0$.

[13] Investment firms choose k and g to maximize profits, $\max_{g,k}\left(q_t g - \iota_t(g) - r_t^k\right)k$. Constant returns to scale imply zero profits in equilibrium, so it doesn't matter who owns them.

Households' Problem.—Households are all identical and have homothetic preferences, so we may consider the problem faced by a representative household. It starts with some wealth w_0 (derived from its initial ownership of capital and government transfers) and chooses a stream of consumption c_h to maximize utility subject to the budget constraint,

$$V_0 = \max_{c_h} U(c_h),$$

subject to

$$E^Q\left[\int_0^\infty e^{-\int_0^t r_m dm} c_{h,t}\, dt\right] \leq w_0.$$

This is equivalent to choosing c_h and the exposure of wealth to aggregate risk σ_w to maximize utility subject to a dynamic budget constraint

$$\frac{dw_t}{w_t} = (r_t + \sigma_{w,t}\pi_t - \tilde{c}_{h,t})dt + \sigma_{w,t} dZ_t$$

and a solvency constraint $w_t \geq 0$, where $\tilde{c}_{h,t} = c_{h,t}/w_t$.[14] Implicit in the second formulation is the fact that since idiosyncratic risks W_i have price zero in equilibrium, it is without loss of generality that households will never choose to be exposed to them.

Intermediaries' Contracts.—Each intermediary would like to borrow from and share risk with the market, but he faces a moral hazard problem with hidden trade: he can secretly steal capital for a private benefit. The contractual environment is developed in detail in the online Appendix. In this section I drop the i subindex to avoid clutter.

Formally, the intermediary starts with net worth $n_0 > 0$ which he gives up in exchange for a full commitment contract $\mathcal{C} = (c, k)$ that specifies his consumption stream c and the capital he will manage k. Both can depend on the history of aggregate shocks Z and his observable return R.[15] Faced with a contract $\mathcal{C}$, the intermediary privately chooses a stealing plan $s \geq 0$, also contingent on the history of Z and R. As a result, the observed return per dollar $q_t k_t$ invested in capital is[16]

$$dR_t = \left(\frac{a - \iota_t(g_t)}{q_t} + g_t + \mu_{q,t} + \sigma_t \sigma'_{q,t} - \tau_t^k - s_t\right)dt + \left(\sigma_t + \sigma_{q,t}\right)dZ_t + \nu_t dW_t.$$

The principal doesn't observe the stealing s, so he doesn't know if bad returns R are due to stealing or just bad luck W.

[14] The link is $w_t = E_t^Q\left[\int_t^\infty e^{-\int_t^u r_m dm} c_{h,u}\, du\right]$ and $\sigma_{w,t} w_t$ is the loading on Z of w_t thus defined.

[15] In principle the contract could also depend on other intermediaries' returns, but this is never optimal.

[16] Notice the total dividend flow from holding capital is the output a plus the rent $r_t^k = q_t g_t - \iota_t(g_t)$.

The intermediary keeps a fraction $\phi_t \in (0,1)$ of the stolen funds $\phi_t q_t k_t s_t$, which can also depend on the history of aggregate shocks Z. He adds them to his consumption (he doesn't have access to hidden savings) so his utility is $U(c + \phi qks)$. Hidden trade is playing a crucial role here, allowing the intermediary to transform stolen capital into consumption goods through markets.[17] Buying and selling capital assets is essential for what intermediaries do, so it is difficult to monitor their activity. Stealing can represent a variety of misbehavior. For example, an intermediary can undersell assets, hurting the principal and benefiting a third party who then shares the spoils (this is the essence of "late trading"). In Section IC I provide concrete examples of the type of intermediary misbehavior that stealing is meant to capture.

In this environment it is always optimal to implement no stealing in equilibrium, $s = 0$.[18] A contract $\mathcal{C} = (c,k)$ is *incentive compatible* if

$$0 \in \arg\max_{s} U(c + \phi qks). \tag{4}$$

Let $\mathbb{IC}$ be the set of incentive compatible contracts. An incentive compatible contract is *optimal* if it minimizes the cost of delivering utility to the agent:

$$J_0(u_0) = \min_{(c,k)\in\mathbb{IC}} E^{Q}\left[\int_0^{\infty} e^{-\int_0^t r_m dm}\left(c_t - q_t k_t \alpha_t\right) dt\right]$$

$$\text{subject to} \quad U(c) \geq u_0,$$

where $\alpha_t \equiv \frac{a - \iota_t(g_t)}{q_t} + g_t + \mu_{q,t} + \sigma_t \sigma'_{q,t} - \tau_t^k - r_t - (\sigma_t + \sigma_{q,t})\pi_t$ is the risk-adjusted expected excess return on capital (determined in equilibrium).[19] We pin down the initial utility with a break even condition. An intermediary with net worth n_0 can buy a contract with cost $J_0(u_0) = n_0$, and get utility u_0. At any point in time t, denote by J_t the continuation cost of the contract.

Implementation with an Equity Constraint.—The optimal contract can be implemented as a constrained portfolio problem with $n_t = J_t$ as the intermediary's net worth. The intermediary raises outside equity e_t and debt d_t to invest in capital: $q_t k_t = n_t + e_t + d_t$. He must keep at least a fraction $\tilde{\phi}_t = \frac{n_t}{n_t + e_t}$ of the total equity for incentive reasons (i.e., "skin in the game"), and the contract specifies his compensation $\tilde{c}_t = c_t / n_t$, but he is free to choose how much to invest in capital k and the exposure to aggregate risk $\sigma_{n,t}$. Since he wants to minimize his exposure to idiosyncratic risk, and he can get aggregate risk in other ways, the retained equity constraint is always binding. Debt yields the risk-free rate r and outside equity

[17] If the intermediary didn't have access to hidden trade, stealing capital wouldn't give him any private benefit, and there wouldn't be a moral hazard problem.

[18] The standard argument applies: if the agent is stealing in equilibrium it's better to just give him what he steals and have him not steal instead. See DeMarzo and Sannikov (2006) or DeMarzo and Fishman (2007) for example.

[19] In the context of investment funds, this is the "gross alpha" on assets. Here in equilibrium intermediaries appropriate all the excess returns, so outside investors only get the market rate of return (zero "net alpha" on equity). See Berk and Green (2004). Other financial institutions also have an implicit "alpha."

yields a return $r + \sigma_{n,t}\pi_t$ and volatility $\sigma_{n,t}dZ_t + \frac{q_t k_t}{n_t + e_t}\nu_t dW_t$. As a result, the intermediary's net worth follows the dynamic budget constraint

$$dn_t = (r_t n_t + q_t k_t \alpha_t - c_t + \sigma_{n,t} n_t \pi_t)dt + \sigma_{n,t} n_t dZ_t + \tilde{\phi}_t q_t k_t \nu_t dW_t. \tag{5}$$

The intermediary chooses k and σ_n to maximize $U(\tilde{c}n)$ subject to (5) and $n_t \geq 0$. Lemma 3 in the online Appendix formalizes implementation and shows this scheme implements the optimal contract.

B. *Competitive Equilibrium*

Take as given the initial capital stock k_0 and the initial distribution of wealth for intermediaries $\{\theta_i > 0\}_{i \in \mathbb{I}}$, such that $\int_{\mathbb{I}} \theta_i di < 1$ (the rest belongs to the representative household).

DEFINITION 1: *A competitive equilibrium is a set of aggregate processes: price of capital q, value of transfers T, risk-free interest rate r, price of aggregate risk π, growth rate g, and the aggregate capital stock k; a contract $\mathcal{C}_i = (c_i, k_i)$ for each intermediary; and a consumption stream c_h for the representative household, such that*

(*i*) *the representative household's consumption and intermediaries' contracts are optimal, with initial wealth $n_{i,0} = \theta_i(q_0 + T_0)k_0$ and $w_0 = (q_0 + T_0)k_0\left(1 - \int_{\mathbb{I}} \theta_i di\right)$;*

(*ii*) *investment is optimal, $\iota_t'(g_t) = q_t$;*

(*iii*) *the value of transfers T satisfies* (3);

(*iv*) *markets clear,*

$$\int_{\mathbb{I}} c_{i,t} di + c_{h,t} = (a - \iota_t(g_t))k_t,$$

$$\int_{\mathbb{I}} k_{i,t} di = k_t;$$

(*v*) *aggregate capital satisfies the law of motion*

$$\frac{dk_t}{k_t} = g_t dt + \sigma_t dZ_t.$$

C. *Discussion of Assumptions*

Financial intermediaries should be interpreted as the insiders who run financial institutions that invest capital on behalf of households, and who must retain an

equity stake for incentive reasons. They could represent different types of financial intermediaries such as hedge funds, private equity (PE) or venture capital (VC) funds, broker/dealers. Commercial banks are sui generis. To the extent that they are involved in extending credit (mortgages, business loans, etc.), they fit into the framework in this paper. However, commercial banks are also in the business of providing liquidity and enjoy deposit insurance or implicit bailouts. This is not included in the model and might require specific regulation (Chari and Kehoe 2013). As is common in the literature, in the model intermediaries hold physical capital, while in reality they hold financial claims on firms and households that actually hold capital. The focus here is on the relationship between these intermediaries and their outside investors, so I abstract from the relationship between intermediaries and final users of capital. In the baseline model I consider only a single type of financial intermediary and homogeneous capital, but a central concern for regulators is how to treat different types of financial institutions and different asset classes. In Section VI, I extend the framework to address these issues.

Equity stakes for insiders are a common financial arrangement designed to align insiders' incentives with outside investors'. He and Krishnamurthy (2013) report an average equity ownership of officers and directors in the finance, insurance, and real estate sectors of 17.4 percent. Hedge funds, PE, and VC typically charge a management fee of 2 percent on assets under management (AUM), plus a "carried interest" of 20 percent of capital gains (above a watermark or hurdle rate).[20] Mutual fund directors and managers usually have their own wealth invested in the fund or bonuses that depend on the fund performance (Chen, Goldstein, and Jiang 2008; Ma, Tang, and Gómez 2015).

The private diversion of funds can represent several forms of misbehavior. Churning, front-running, market timing, late trading, and bid-ask spread manipulation are important concerns for mutual and hedge funds. PE and VC funds may overpay or undersell firms, and they also charge large and obscure portfolio company fees. They all involve benefiting some privileged investors, insiders, or third parties at the cost of investors, usually by letting them trade at stale prices or with inside information, in exchange for a quid pro quo (for example, fee-generating "sticky assets").[21]

[20] Notice even the 2 percent already embodies an equity stake, because good returns increase AUM and therefore fees. The 20 percent is designed to provide strong incentives. Ackermann, McEnally, and Ravenscraft (1999) finds that this does indeed improve performance, raising the Sharpe ratio by 0.15.

[21] The 2003 mutual fund scandal is a salient example. Bank of America's (BoA) Nations Fund allegedly allowed late trading by privileged investors. Mutual funds shares are priced at 4 PM, but these privileged investors were allowed to buy after 4 PM at those stale prices. This allowed them to buy fund shares when the value of the assets was higher than the price, hurting the other investors. They then shared the spoils with the mutual fund management by depositing "sticky assets" in fee-generating funds from the same family. See Zitzewitz (2003) for a more detailed explanation of late trading and an empirical analysis of its relevance.

The same scandal involved accusations of front-running (allowing privileged clients or partners to benefit from the price impact of large movements in the fund portfolio) that led to the resignation of the chairman of Strong Mutual Funds. Fund families may also favor some funds at the expense of others; see Gaspar, Massa, and Matos (2006). Lack (2012) describes how hedge funds may manipulate the bid-ask spread.

Portfolio-company fees account for transactions, advisory and monitoring, consulting, etc. They are levied directly on the acquired firms, so the fund investors never see them. They only notice lower returns on their investment. In addition, PE and VC funds have incentives to postpone shutting down worthless investments to continue collecting management fees. See Ang (2014, pp. 610–12) and Phalippou (2009).

Hidden trade is essential in all these examples. Since financial intermediaries' job is to buy and sell capital assets, it is difficult to determine if they are doing the right trades or at the right prices.[22] In the model, intermediaries steal capital and trade it for consumption goods. If they couldn't trade capital there wouldn't be a moral hazard problem (they don't value capital by itself, only consumption). Contrast this with a factory worker who may also use or manage very valuable capital, like a forklift. He too may have a moral hazard problem that may impact the value of the capital. But since their job does not involve buying and selling capital, there is no hidden trade problem. Note that while intermediaries ability to trade capital for consumption goods is very important, the presence of spot markets versus futures markets is not. Everything in this model fits into an Arrow-Debreu intertemporal framework.

I allow agents to write complete, long-term contracts. It is important that we don't impose any ad hoc constraints on private contracts in order to understand the efficiency properties of the competitive equilibrium. In particular, contracts can be made contingent on all observable variables, including all aggregate shocks. This is also realistic in practice: financial markets provide ample opportunities for intermediaries to hedge aggregate risk. Long-term contracts make sense when thinking of contracts between financial institutions and their insiders. In addition, they provide a clean contractual environment which makes comparisons with the social planner's allocation straightforward. As it turns out, optimal long-term contracts are renegotiation proof, and the distinction between short- and long-term contracts is important for the efficiency of the competitive equilibrium.

Finally, an intermediary's idiosyncratic risk W_i represents the risk associated with their specific investment activity. For example, if two hedge funds invest a dollar in stocks, they will obtain different returns depending on their specific trading strategy. The idiosyncratic risk $W_{i,t}$ is not the risk in each of the assets they buy, since they may diversify or hedge, but rather the idiosyncratic risk in their overall investment strategy which reflects their skill, information, or luck. Likewise, PE or VC funds must pick some firms/startups to invest in based on their business analysis or private information, and are therefore exposed to idiosyncratic risk. Commercial banks that issue mortgages and business loans are very diversified with respect to the risk of each of those loans. But they are exposed to idiosyncratic risk that reflects their business strategy: e.g., some banks may have a large exposure to the Miami real estate market, others to the auto industry, or they may differ in their securitization strategies, etc., so that two banks will generally get different returns.

[22] It may seem surprising then that in the model k is contractible. While the contract knows that the intermediary has assets worth $q_t k_t$, it can't tell exactly what trades the intermediary is doing, or if these trades are the right ones. In addition, as it turns out, under the optimal contract the intermediary has no incentives to deviate on the k margin, so there is no need to actually monitor him there (see Lemma 3 in the online Appendix).

II. Solving the Competitive Equilibrium

A. *Recursive Formulation*

Optimal Contracts.—Intermediaries' optimal contracts are recursive in their continuation utility $U_{i,t}$. Drop the i subscript to simplify notation. We can use Lemma 1 in the online Appendix to write the law of motion of an intermediary's continuation utility

$$dU_t = -f(c_t, U_t)dt + \sigma_{U,t}dZ_t + \tilde{\sigma}_{U,t}dW_t. \tag{6}$$

This can be interpreted as a promise keeping constraint. If the intermediary has been promised utility U, this must be delivered by consumption c, either today or in the future. His continuation utility can be exposed to both aggregate risk Z and idiosyncratic risk W. Exposing the intermediary to idiosyncratic risk W is costly because he is risk averse and the market doesn't price idiosyncratic risk (the first best has full insurance against idiosyncratic risk), but it is necessary for incentive reasons. If the intermediary steals, he adds ϕqks to his consumption, but bad returns are more likely to be observed, conditional on the observable aggregate shock. To deter him from stealing, the optimal contract must give him lower continuation utility after bad outcomes are observed. Lemma 2 in the online Appendix shows that, for the parameter values $\gamma > 1$ and $\psi > 1$, the contract $\mathcal{C} = (c, k)$ is incentive compatible if and only if

$$0 \in \arg\max_{s \geq 0} f(c_t + \phi_t q_t k_t s, U_t) - \tilde{\sigma}_{U,t}\frac{s}{\nu_t} - f(c_t, U_t).$$

Taking FOC we obtain

$$\tilde{\sigma}_{U,t} \geq \partial_c f(c_t, U_t)\phi_t q_t k_t \nu_t = \frac{c_t^{-1/\psi}}{((1-\gamma)U_t)^{\frac{\gamma-1/\psi}{1-\gamma}}}\phi_t q_t k_t \nu_t \geq 0. \tag{7}$$

The incentive compatible (IC) constraint will be binding in the optimal contract. We can also verify that if contract $\mathcal{C} = (c, k)$ is incentive compatible, so is a scaled up version of it $\mathcal{C}' = (\kappa c, \kappa k)$. In consequence, the cost function of the optimal contract takes the form

$$J_t = \xi_t((1-\gamma)U_t)^{\frac{1}{1-\gamma}}. \tag{8}$$

Thanks to homothetic preferences, the optimal contract is linear in the utility of the intermediary measured in consumption units $x_t = ((1-\gamma)U_t)^{\frac{1}{1-\gamma}}$ (up to a constant). The endogenous process ξ captures the stochastic investment opportunities facing

the intermediary, and tells us what is the cost of delivering utility x to him. It depends only on the history of aggregate shocks Z and follows the law of motion,

$$\frac{d\xi_t}{\xi_t} = \mu_{\xi,t}dt + \sigma_{\xi,t}dZ_t, \tag{9}$$

which must be determined in equilibrium. The Hamilton-Jacobi-Bellman (HJB) equation associated with the optimal contract is

$$r_t J_t dt = \min_{c,k,\sigma_U} (c - q_t k \alpha_t)\, dt + E_t^Q[dJ_t]. \tag{10}$$

We can use Ito's lemma together with (6) and (9) to expand the last term. Since expectations are taken under the equivalent martingale measure Q, it is useful to write $Z_t = Z_t^Q - \int_0^t \pi_u du$, where Z^Q is a Brownian motion under Q. We can normalize the controls $c_t = \hat{c}_t x_t$, $k_t = \hat{k}_t x_t$, and $\sigma_{U,t} = \sigma_{x,t}(1-\gamma)\, U_t$. Intermediaries get consumption and capital proportional to their continuation utility measured in consumption units x, i.e., they all get the same $\hat{c}$, $\hat{k}$, and σ_x. The HJB equation then takes the following form:[23]

$$\begin{aligned} r_t \xi_t = \min_{\hat{c},\hat{k},\sigma_x} \hat{c} - q_t \hat{k} \alpha_t + \xi_t \Bigg\{ & \frac{1}{1-\frac{1}{\psi}} \left(\rho - \hat{c}^{1-1/\psi}\right) - \sigma_x \pi_t + \mu_{\xi,t} - \sigma_{\xi,t}\pi_t \\ & + \frac{1}{2}\gamma\sigma_x^2 + \frac{1}{2}\gamma\left(\hat{c}^{-1/\psi}\phi_t q_t \hat{k} \nu_t\right)^2 + \sigma_{\xi,t}\sigma_x \Bigg\}. \end{aligned} \tag{11}$$

It is easy to see from (10) that the intermediary's net worth $n_t = J_t = \xi_t x_t$ follows the law of motion (5). In fact, Lemma 3 in the online Appendix shows that the optimal contract can be implemented with a constrained portfolio problem. The intermediary must retain an equity stake $\frac{n_t}{n_t + e_t} = \tilde{\phi}_t$, and his compensation out of his net worth is specified by $\tilde{c}_t = c_t / n_t$. Both depend on the history of returns R and aggregate shocks Z. The intermediary is free to choose how much to invest in capital k and his exposure to aggregate shocks σ_n. The retained equity stake $\tilde{\phi}_t$ is given by

$$\tilde{\phi}_t = \xi_t \hat{c}_t^{-1/\psi} \phi_t = \xi_t^{1-1/\psi} \left(\frac{c_t}{n_t}\right)^{-1/\psi} \phi_t. \tag{12}$$

[23] Existence of the optimal contract requires $\psi > 2$. This is related to the assumption of no hidden savings. Once we assume this, the objective function in the HJB is convex, and the FOC are sufficient for optimality. The online Appendix provides a verification theorem.

The intermediary's equity stake provides incentives to deter misbehavior, but forces him to keep a fraction $\tilde{\phi}_t$ of the idiosyncratic risk from his capital. Note that with long-term contracts the principal can relax the equity constraint, $\tilde{\phi}_t < \phi_t$, and improve the idiosyncratic risk sharing problem by increasing the agent's consumption. A large $\hat{c}_t$ reduces the private benefit of stealing, as shown in the IC constraint (7). This allows private contracts some flexibility on the equity constraint, and is also the reason why the intermediary cannot be allowed to choose his compensation on his own.[24]

Finally, notice that the optimal contract is renegotiation-proof. After any history, the continuation contract is the cheapest way of delivering the promised utility to the intermediary. If there was a cheaper incentive compatible contract that delivered at least as much utility, we could scale it down to deliver the same utility, at an even lower cost. And because it delivers the same utility, it would not affect incentives. Lemma 4 in the online Appendix formalizes this result. The counterpart of this result is that the planner's allocation will also be renegotiation-proof, as we'll see in Section III.

Households' Problem.—Households have a value function,

$$V_t(w_t) = \frac{(\zeta_t w_t)^{1-\gamma}}{1-\gamma},$$

where ζ captures their endogenously stochastic investment possibilities. It depends only on the history of aggregate shocks Z with law of motion,

$$\frac{d\zeta_t}{\zeta_t} = \mu_{\zeta,t}dt + \sigma_{\zeta,t}dZ_t, \tag{13}$$

which we must find in equilibrium. Analogously to intermediaries, we can interpret ζ^{-1} as the cost of delivering utility $h_t = ((1-\gamma)V_t)^{\frac{1}{1-\gamma}}$ measured in consumption units to households.

After some algebra, the associated HJB equation for ζ is

$$\begin{aligned}\frac{\rho}{1-1/\psi} = \max_{\tilde{c}_h, \sigma_w} & \frac{\tilde{c}_h^{1-1/\psi}}{1-1/\psi}\zeta_t^{1/\psi-1} + r_t + \sigma_w\pi_t \\ & - \tilde{c}_h + \mu_{\zeta,t} - \frac{\gamma}{2}\sigma_{\zeta,t}^2 - \frac{\gamma}{2}\sigma_w^2 + (1-\gamma)\sigma_{\zeta,t}\sigma_w.\end{aligned} \tag{14}$$

[24] In the setting with short-term contracts in Di Tella (2017), we get the same characterization with $\tilde{\phi}_t = \phi_t$ because contracts cannot control the intermediary's consumption. If we further assume that aggregate risk cannot be traded, we get the contractual setting in Brunnermeier and Sannikov (2014) or He and Krishnamurthy (2012), where $\sigma_{n,t}n_t = q_t k_t \phi_t(\sigma_t + \sigma_{q,t})$.

Value of Transfers.—Since T only depends on the history of aggregate shocks Z, we can write

$$\frac{dT_t}{T_t} = \mu_{T,t}dt + \sigma_{T,t}dZ_t. \tag{15}$$

In equilibrium T must satisfy a no-arbitrage pricing equation

$$\frac{q_t\tau_t^k}{T_t} + \mu_{T,t} + g_t + \sigma_t\sigma_{T,t}' - r_t = (\sigma_{T,t} + \sigma_t)\pi_t. \tag{16}$$

Recursive Equilibrium.—Since contracts are linear in $x_{i,t}$, and the economy is scale invariant to the level of capital k_t, we can use

$$X_t = \frac{\int_{\mathbb{I}} x_{i,t}di}{k_t}$$

as an endogenous aggregate state variable. Here, X_t captures the aggregate utility promised to intermediaries (normalized by k_t), and is the same endogenous aggregate state variable that the planner's problem will use, which makes comparisons straightforward. Market clearing conditions can be written as follows:

$$\tilde{c}_h(q_t + T_t - \xi_t X_t) + \hat{c}_t X_t = a - \iota_t(g_t), \quad [\textit{consumption goods}] \tag{17}$$

$$\hat{k}_t = \frac{1}{X_t}. \quad [\textit{capital}] \tag{18}$$

Using Ito's lemma and the equilibrium conditions $\hat{k}_t X_t = 1$ and $q_t = \iota_t'(g_t)$, we get a law of motion for X_t,

$$\frac{dX_t}{X_t} = \mu_{X,t}dt + \sigma_{X,t}dZ_t, \tag{19}$$

$$\mu_{X,t} = \frac{\rho}{1-1/\psi} - \frac{\hat{c}_t^{1-1/\psi}}{1-1/\psi} + \frac{\gamma}{2}\sigma_{x,t}^2 + \frac{\gamma}{2}\left(\hat{c}_t^{-1/\psi}\phi_t\frac{\iota_t'(g_t)\nu_t}{X_t}\right)^2 - g_t - \sigma_t\sigma_{x,t} + \sigma_t^2,$$

$$\sigma_{X,t} = \sigma_{x,t} - \sigma_t.$$

We look for an equilibrium with state variables X_t and Y_t, where equilibrium objects such as q, T, ξ, ζ are functions of (X, Y). Recall that several features of the environment, such as ν_t, σ_t, $\iota_t(g)$, or ϕ_t are already functions of the exogenous state variable Y_t. In addition, for this to work it must be the case that taxes also share the same state variables, i.e., $\tau_t^k = \tau^k(X_t, Y_t)$. This will be the case both for the unregulated competitive equilibrium and for the implementation of the optimal allocation. We can then use Ito's lemma to transform equilibrium conditions into a system of second-order PDEs. The online Appendix shows the solution method in detail.

B. *Asset Prices and Financial Amplification Channel*

Intermediaries' FOC with respect to $\hat{k}$, together with $n_t = \xi_t x_t$, and the market clearing condition for capital gives us a pricing equation for capital

$$\underbrace{\frac{a-\iota_t(g_t)}{q_t} + g_t + \mu_{q,t} + \sigma_t\sigma'_{q,t} - \left(r_t + \tau_t^k\right) - \left(\sigma_t + \sigma_{q,t}\right)\pi_t}_{\text{risk-adjusted expected excess return } \equiv\ \alpha_t} = \underbrace{\gamma\frac{q_t k_t}{n_t}\left(\tilde{\phi}_t\nu_t\right)^2}_{\text{id. risk premium}}. \tag{20}$$

Although the market price of idiosyncratic risks $\{W_i\}$ is zero (they can be aggregated away), capital must pay a premium for this risk. Because of the moral hazard problem, intermediaries must be exposed to idiosyncratic risk proportionally to the value of capital they manage, as in (7). This is costly because intermediaries are risk averse, so they will demand a premium for holding capital.[25] A large excess return α_t implies a low price of capital, and therefore low investment and growth through $\iota'_t(g_t) = q_t$.

Financial conditions affect the economy through equation (20). During periods of instability with high idiosyncratic risk ν_t, or when financial frictions are tight (high $\tilde{\phi}_t$), capital must yield a large excess return α_t. In addition, if after a bad aggregate shock financial losses are concentrated on the balance sheets of intermediaries (n_t falls proportionally more than $q_t k_t$), intermediaries will require an even larger excess return on capital α_t. This creates a financial amplification channel that further depresses investment and asset prices.

But why would financial losses be concentrated on the balance sheets of intermediaries? Since agents are perfectly free to share aggregate risk, the market equalizes the ratio of marginal utility of wealth between intermediaries and households, i.e., the volatility of $\frac{\xi_t^{\gamma-1} n_t^{-\gamma}}{\zeta_t^{1-\gamma} w_t^{-\gamma}}$ is zero. To understand what this implies about how financial gains or losses are shared between households and intermediaries, consider the marginal rate of substitution between intermediaries' and households' utilities (measured in consumption units),

$$\Lambda_t = \xi_t\zeta_t,$$

with endogenous law of motion $d\Lambda_t/\Lambda_t = \mu_{\Lambda,t}dt + \sigma_{\Lambda,t}dZ_t$. If we increase an intermediary's utility by Δx, the cost of his contract increases by $\xi_t \Delta x$. With this money a household could obtain utility $\Delta h_t = \xi_t\zeta_t\Delta x$. The MRS Λ_t is therefore the cost of intermediaries' utility in terms of foregone utility by households, both measured in consumption units. The only difference between intermediaries and households is that intermediaries can obtain an excess return α_t by investing in capital and taking on idiosyncratic risk. The cost of intermediaries' utility Λ_t is therefore low when they expect large excess returns on capital α_t looking forward.

[25] As we would expect, the premium on idiosyncratic risk vanishes if there is no moral hazard ($\phi_t = 0$, which implies $\tilde{\phi}_t = 0$) or no idiosyncratic risk ($\nu_t = 0$).

We can now use the FOCs for σ_x and σ_w, as well as $n_t = \xi_t x_t$ to obtain an expression for the exposure to aggregate risk of intermediaries relative to households:

$$\sigma_{n,t} - \sigma_{w,t} = \frac{\gamma - 1}{\gamma} \sigma_{\Lambda,t}. \tag{21}$$

Here we have two opposing effects. On the one hand, there is a *substitution effect*: intermediaries should have more net worth when the cost of intermediaries' utility Λ_t is low in order to get more "bang for the buck." But there is also an *income effect*: intermediaries need more net worth in order to achieve any given utility level when the cost of their utility Λ_t is high. *In the empirically relevant case with risk aversion γ greater than 1, financial losses are concentrated on intermediaries' balance sheets after aggregate shocks that reduce the cost of providing utility to them Λ.* Essentially, intermediaries are willing to take large financial losses up front if they expect large excess returns α_t looking forward (so Λ_t is low). This is the same mechanism as in Di Tella (2017), but for general aggregate shocks.

We can also think of aggregate risk sharing in terms of utility. Using $n_t = \xi_t x_t$, and $w_t = \zeta_t^{-1} h_t$, we obtain

$$\sigma_{x,t} - \sigma_{h,t} = -\frac{1}{\gamma} \sigma_{\Lambda,t}. \tag{22}$$

Utility losses are concentrated on households after aggregate shocks that reduce the cost of intermediaries' utility Λ_t. This means that for the relevant case with $\gamma > 1$, if financial losses are disproportionally concentrated on the balance sheets of intermediaries, e.g., during a financial crisis, it is households who suffer disproportionate losses in utility terms.

Of course, α_t and Λ_t are endogenous equilibrium objects whose behavior depends on the type of aggregate shocks hitting the economy. In Section V, I show that uncertainty shocks that raise idiosyncratic risk ν_t create a financial amplification channel with financial losses concentrated on the balance sheets of intermediaries. In general, the economy might be hit by several different types of aggregate shocks at the same time. Equations (21) and (22) allow us to understand why aggregate risk may be concentrated even if there are no limits to aggregate risk sharing, and to study the welfare and policy implications for arbitrary aggregate shocks.

Understanding the MRS Λ_t Better.—Since Λ plays an important role in the allocation of aggregate risk, it is worth studying in more detail. Λ correctly measures the marginal rate of substitution (MRS) between intermediaries' and households' utility (in consumption units), taking into account that giving consumption to intermediaries helps relax the IC constraint (7). To understand this better, look at the FOC for intermediaries' consumption $\hat{c}$,

$$\xi_t \hat{c}_t^{-1/\psi} + \underbrace{\xi_t \frac{\gamma}{\psi} \left(\phi q_t \hat{k}_t \nu_t\right)^2 \hat{c}_t^{-2/\psi - 1}}_{\text{relax IC}} = 1. \tag{23}$$

The first term captures the standard intertemporal trade-off, while the second term captures the fact that by giving the intermediary more consumption, we can reduce the marginal utility of consumption and therefore make stealing less attractive, as shown in the IC constraint (7). This is reflected in the expression for the intermediary's equity stake $\tilde{\phi}_t = \xi_t \hat{c}_t^{-1/\psi} \phi_t < \phi_t$. If we ignored the second term, we would get $\tilde{\phi}_t = \phi_t$. In contrast, the FOC for households' consumption,

$$\tilde{c}_{h,t}^{-1/\psi} \zeta_t^{1/\psi - 1} = 1, \tag{24}$$

features only the standard consumption smoothing trade-off. Putting the FOC for $\hat{c}$ and $\tilde{c}_h$ together, and using $\hat{c}_{h,t} = c_t / h_t$ (analogous to $\hat{c}_t = c_t / x_t$ for intermediaries), we obtain

$$\Lambda_t = \frac{\hat{c}_{h,t}^{-1/\psi}}{\hat{c}_t^{-1/\psi} + \frac{\gamma}{\psi} \left(\phi_t \iota_t'(g_t) \frac{\nu_t}{X_t} \right)^2 \hat{c}_t^{-2/\psi - 1}}. \tag{25}$$

The numerator is the marginal utility of households' consumption (in consumption units), $\hat{c}_h^{-1/\psi} = \partial_{c_h} h_t$. The denominator captures the marginal utility for intermediaries consumption $\hat{c}^{-1/\psi}$ plus the benefit of relaxing the IC constraint. This implies that while the ratio of marginal utility of wealth $\frac{\xi_t^{\gamma-1} n_t^{-\gamma}}{\zeta_t^{1-\gamma} w_t^{-\gamma}}$ is equalized across aggregate states, this is not true for the ratio of marginal utility of consumption $\frac{\partial_c f(c_t, U_t)}{\partial_c f(c_{h,t}, V_t)}$, as would be the case in a standard model with complete markets. This is because incentives to distort the consumption smoothing margin to improve idiosyncratic risk sharing depend on the aggregate state of the economy (if they were invariant to aggregate shocks, $\frac{\partial_c f(c_t, U_t)}{\partial_c f(c_{h,t}, V_t)}$ would actually be equalized across aggregate states). This feature of long term contracts plays an important role in the welfare analysis, since private contracts internalize the relative value of relaxing the IC constraints across aggregate states.

III. Planner's Problem

In this section I characterize the best allocation that can be achieved by a social planner who faces the same informational frictions as private agents. Hidden trade creates an externality: intermediaries don't internalize that by demanding capital and bidding up its price, they force others to bear more idiosyncratic risk. The socially optimal allocation can be implemented by a tax on assets.

A. *Setting*

Consider a social planner who faces the same informational frictions as private agents in the market. He can (i) control households' consumption; (ii) give consumption and capital to intermediaries to manage, but they can secretly divert it; (iii) give capital and consumption goods to investment firms and order them to

deliver a flow of new capital. As in the competitive equilibrium, intermediaries and firms have access to hidden trade in capital.[26] An intermediary with a flow of stolen capital $k_{i,t}s_{i,t}$ can sell it to a firm at a competitive black market price $\tilde{q}_t$. The firm will produce less new capital and present the stolen capital to the planner, so that the hidden trade is not detected.

To formalize this, consider an investment firm that receives an order to use k_t units of capital and $\iota_t(g_t)k_t$ consumption goods to deliver a flow of new capital $g_t k_t$. It can instead buy a flow of stolen capital $k_t s$ and do actual investment $k_t \iota_t(\tilde{g})$ in order to maximize its surplus consumption (that it rebates to its owners)

$$\max_{s,\tilde{g}} \iota_t(g_t)k_t - \iota_t(\tilde{g}_t)k_t - \tilde{q}_t s k_t,$$

subject to

$$(\tilde{g} + s)k_t = g_t k_t.$$

Optimality implies $\tilde{q}_t = \iota_t'(\tilde{g})$, so to implement investment rate g_t and no stealing in equilibrium, $s_t = 0$, the black market price of capital must be

$$\tilde{q}_t = \iota_t'(g_t). \tag{26}$$

Notice this is precisely the equilibrium price of capital in the competitive equilibrium, which is consistent with this environment.

A plan $\mathcal{P} = \left(c_h, g, k, \{c_i, k_i\}_{i\in[0,1]}\right)$ is a consumption stream for the representative household c_h and a growth rate g and aggregate capital k, which can depend on the history of aggregate shocks Z; and consumption and capital (c_i, k_i) for each intermediary i, which can depend also on his history of idiosyncratic outcomes. Faced with a feasible plan $\mathcal{P}$, each intermediary chooses a stealing plan s_i and gets consumption $\tilde{c}_i = c_i + \phi\tilde{q}k_i s_i$. As in the private problem, it is optimal to implement no stealing always, $s_i = 0$. With this in mind, we say a plan $\mathcal{P}$ is *feasible* if it satisfies the aggregate consistency conditions

$$c_{h,t} + \int_{\mathbb{I}} c_{i,t}di = (a - \iota_t(g_t))k_t, \tag{27}$$

$$\int_{\mathbb{I}} k_{i,t} = k_t, \tag{28}$$

and aggregate capital follows the law of motion

$$dk_t = g_t k_t dt + \sigma_t k_t dZ_t. \tag{29}$$

[26] See Farhi, Golosov, and Tsyvinski (2009) or Kehoe and Levine (1993).

A feasible plan is *incentive compatible* if choosing $s_i = 0$ is optimal for every intermediary:

$$0 \in \arg\max_s U(c_i + \phi \iota'(g) k_i s). \tag{30}$$

This IC constraint is the same as the IC in private contracts (4), except that the planner internalizes that by controlling g he can relax the moral hazard problem. This is the source of inefficiency in this model. Let $\mathbb{ICP}$ be the set of incentive compatible plans. Given initial utility levels for each intermediary $\{u_i^0\}_{i\in[0,1]}$, an incentive compatible plan $\mathcal{P}$ is *optimal* if it maximizes households' utility subject to delivering utility u_i^0 to each intermediary:

$$\max_{\mathcal{P}\in\mathbb{ICP}} U(c_h),$$

subject to

$$U_i(c_i) = u_i^0.$$

B. *A Recursive Formulation of the Planner's Problem*

Just as in the competitive equilibrium, we look for an optimal mechanism that is recursive in the continuation utility of intermediaries $\{U_i\}_{i\in\mathbb{I}}$ and the aggregate state variables. Each intermediary's utility still follows the law of motion given by (6), and the IC constraint is like (7) with q_t replaced by $\tilde{q}_t = \iota_t'(g_t)$:

$$\tilde{\sigma}_{U,i,t} \geq f_c(c_{i,t}, U_{i,t})\phi_t \iota_t'(g_t) k_{i,t} \nu_t = \frac{c_{i,t}^{-1/\psi}}{((1-\gamma)U_{i,t})^{\frac{\gamma-1/\psi}{1-\gamma}}} \phi_t \iota_t'(g_t) k_{i,t} \nu_t \geq 0, \tag{31}$$

and it will be binding in the optimal allocation.

Introduce $x_{i,t} = \left((1-\gamma)U_{i,t}\right)^{\frac{1}{1-\gamma}}$ as in the private contract, with $c_{i,t} = \hat{c}_{i,t} x_{i,t}$ and $k_{i,t} = \hat{k}_{i,t} x_{i,t}$, and $\sigma_{U,i,t} = \sigma_{x,i,t}(1-\gamma)U_{i,t}$. We can verify that, just as in the private contract, the planner will choose the same $\hat{c}_{i,t} = \hat{c}_t$, $\hat{k}_{i,t} = \hat{k}_t$, and $\sigma_{x,i,t} = \sigma_{x,t}$ for all intermediaries. The planner's problem must be recursive in the same endogenous state variable as the competitive equilibrium $X_t = \int_{\mathbb{I}} x_{i,t} di / k_t$ which captures the aggregate continuation utility of intermediaries, and the exogenous state variable Y. From the consistency conditions (27) and (28) we obtain $c_{h,t} = (a - \iota_t(g_t) - \hat{c}_t X_t) k_t$ and $\tilde{k}_t = X_t^{-1}$. Thanks to homothetic preferences and the linear technology, the planner's value at time t then takes the following power form:

$$\frac{(S_t k_t)^{1-\gamma}}{1-\gamma}$$

for some process S_t which depends only on the history of aggregate shocks Z. The variable S_t captures the planner's value (households' utility) in consumption units, normalized by capital: $S_t = ((1-\gamma)V_t)^{\frac{1}{1-\gamma}}/k_t = h_t/k_t$. It is analogous to intermediaries' continuation utility $X_t = \int_{\mathbb{I}} x_{i,t}\, di/k_t$. Likewise, define $\hat{c}_{h,t} = c_{h,t}/h_t$, analogous to $\hat{c}_t = c_t/x_t$ for intermediaries. We look for a value function S and the policy functions $\hat{c}$, g, and σ_x, all functions of (X, Y), and the law of motion of X is given by (19). The HJB equation associated to the planner's problem is

$$\frac{\rho}{1-1/\psi} = \max_{g,\hat{c},\sigma_x} \frac{(a-\iota(g)-\hat{c}X)^{1-1/\psi}}{1-1/\psi} S^{1/\psi-1} + \mu_S + g - \frac{\gamma}{2}\sigma_S^2 - \frac{\gamma}{2}\sigma^2 + (1-\gamma)\sigma_S\sigma, \tag{32}$$

where μ_S and σ_S are obtained from Ito's lemma on $S(X, Y)$. The planner's problem boils down to solving a second-order PDE for $S(X, Y)$. The online Appendix describes the procedure in detail.

The planner's allocation is renegotiation-proof. The only way of delivering more utility to intermediaries is to reduce the utility of the household, conditional on the exogenous aggregate state Y. In other words, $S'_X(X, Y) < 0$ for all (X, Y). Intuitively, the planner can always reduce intermediaries' consumption and increase households'. This gives more utility to households (the planner's objective function), and increases intermediaries' promised utility X (see (19)). If $S'_X(X, Y) > 0$ the planner would benefit from this deviation.[27] This is the planner's counterpart of the renegotiation-proofness of private contracts.

C. *Externality*

Hidden trade creates an externality in the competitive equilibrium. Because intermediaries cannot be prevented from trading capital, the private benefit of stealing depends on the value of capital, $q_t = \iota'_t(g_t)$. The social planner is willing to give up investment/growth g_t in order to reduce the private benefit of stealing and therefore relax the constraints on idiosyncratic risk sharing. This is reflected in the FOC for g:

$$\underbrace{\hat{c}_{h,t}^{-1/\psi}}_{\partial_{c_h}(Sk)} \iota'_t(g_t)(1+\eta_t) = \underbrace{S_t + \Lambda_t X_t}_{\partial_k(Sk)} \tag{33}$$

with

$$\eta_t = \frac{\Lambda_t X_t}{\hat{c}_{h,t}^{-1/\psi}} \gamma \left(\hat{c}_t^{-1/\psi}\phi_t \frac{\nu_t}{X_t}\right)^2 \iota''_t(g). \tag{34}$$

[27] As we'll see below, the FOC (37) establishes $S'_X = -\Lambda = \frac{\hat{c}_{h,t}^{-1/\psi}}{\hat{c}_t^{-1/\psi} + \frac{\gamma}{\psi}\left(\phi_t \iota'_t(g_t)\frac{\nu_t}{X_t}\right)^2 \hat{c}_t^{-2/\psi-1}} > 0.$

Here $\Lambda = -S'_X$ is the MRS between intermediaries' and households' utility (recall in the competitive equilibrium we had $\Lambda = \xi\zeta$). The right-hand side of (33) captures the marginal benefit of having more capital, as we would expect. The first term on the left-hand side captures the marginal utility cost of reducing households' consumption to increase investment. The second term η_t captures the externality. Higher investment increases the marginal cost of capital $\iota'_t(g_t)$ and therefore raises the private benefit of stealing. The planner realizes that if he wants to raise investment, he must expose intermediaries to more idiosyncratic risk, as in (31). Since they must be compensated for the risk, households need to give up more consumption. As a result, the actual marginal cost of producing more capital (in consumption units) is $\iota'_t(g_t)(1+\eta_t)$.

Private agents in the competitive equilibrium don't internalize this trade-off between investment and idiosyncratic risk sharing. They don't realize that when they demand capital and bid up its price, they create a moral hazard problem for everyone else. We can obtain an analogous equation for the competitive equilibrium from $\iota'_t(g_t) = q_t$, $S_t = \zeta_t(q_t + T_t - \xi_t X_t)$, and the FOC for $\tilde{c}_h$:

$$\hat{c}_{h,t}^{-1/\psi}\iota'_t(g_t)\left(1+\frac{T_t}{q_t}\right) = S_t + \Lambda_t X_t. \tag{35}$$

The right-hand side $S_t + \Lambda_t X_t = \hat{c}_{h,t}^{-1/\psi}(q_t + T_t)$ correctly measures the marginal value of capital in the competitive equilibrium allocation[28] $\partial_k(Sk)$, and $\hat{c}_{h_t}^{-1/\psi}$ measures the marginal utility of households' consumption $\partial_{c_h}(Sk)$, so efficiency requires $T_t/q_t = \eta_t$. Notice that if the function $\iota_t(g_t)$ was linear, the planner wouldn't be able to affect the price of capital by distorting g_t, so there would be no externality, $\eta_t = 0$.[29]

The externality captured in equation (33) is the only source of inefficiency. The FOC for σ_x yields, after some algebra,

$$\sigma_{x,t} - \sigma_{h,t} = -\frac{1}{\gamma}\sigma_{\Lambda,t}, \tag{36}$$

as in the competitive equilibrium, and from the FOC for $\hat{c}$ we obtain[30]

$$\Lambda_t = \frac{\hat{c}_{h,t}^{-1/\psi}}{\hat{c}_t^{-1/\psi} + \frac{\gamma}{\psi}\left(\phi\,\iota'_t(g_t)\frac{\nu_t}{X_t}\right)^2\hat{c}_t^{-2/\psi-1}} > 0 \tag{37}$$

[28] If we integrate (20) and use (3), we get

$$q_t + T_t = E_t^Q\left[\int_t^\infty e^{-\int_t^s r_u du}\left(a - \iota_s(g_s) - \xi_s\gamma\left(\hat{c}_s^{-1/\psi}\phi_s\nu_s q_s\right)^2\hat{k}_s\right)e^{\int_t^s\left(g_u-\frac{1}{2}\sigma_u^2\right)du+\int_t^s\sigma_u^2 dZ_u}\right].$$

Suppose we get an extra unit of capital and want to keep the whole process for g and x unchanged. The extra capital produces consumption net of investment $a - \iota_s(g_s)$. However, since intermediaries hold more capital, they must be exposed to more idiosyncratic risk. To keep x unchanged, we need to give them more consumption $\xi_s\gamma\left(\hat{c}_s^{-1/\psi}\phi_s\nu_s q_s\right)^2\hat{k}_s$, where ξ_s already takes into account that the extra consumption also helps to partially offset the increase in idiosyncratic risk. What remains can be added to households' consumption, and the discounted expectation under Q correctly evaluates it in terms of current consumption.

[29] The intuition is the same as in Kehoe and Levine (1993).

[30] $\Lambda_t > 0$ implies $S'_X < 0$, so the planner's optimal allocation is renegotiation-proof: we can only give more utility to households by giving less utility to intermediaries.

which coincides with equation (25) in the competitive equilibrium. Private contracts internalize that giving intermediaries more consumption relaxes the equity constraint and improves idiosyncratic risk sharing. This means that the debt/equity margin is efficient, and private contracts evaluate the MRS between intermediaries' and households' utility Λ correctly. As a result, the allocation of aggregate risk is efficient: if we fix the process for investment in the competitive equilibrium and we allow the planner to only modify consumption $\hat{c}$ and the allocation of aggregate risk σ_x, he would choose not to. This is an important feature of the competitive equilibrium with long-term contracts. In models with incomplete aggregate risk sharing, the MRS is not equalized across aggregate states, and the allocation of aggregate risk is inefficient.[31] Even with complete aggregate risk sharing, short-term contracts, such as in Di Tella (2017), don't internalize how giving intermediaries more wealth/consumption can improve idiosyncratic risk sharing, so they evaluate the MRS between intermediaries' and households' utility Λ as the ratio of marginal utility of consumption $\hat{c}_{h,t}^{-1/\psi}/\hat{c}_t^{-1/\psi}$. If the wedge between this and the correct MRS is correlated with aggregate shocks, the allocation of aggregate risk is inefficient and the competitive equilibrium can be improved by a planner who only regulates the allocation of aggregate risk.[32]

Of course, if we compare the competitive equilibrium with the planner's optimal allocation, we will in general find different MRS Λ, simply because the allocations are different, i.e., for the same states (X, Y) there is a wedge $\Lambda^{SP}/\Lambda^{CE} \neq 1$. If this wedge is correlated with aggregate shocks, the allocation of aggregate risk in the competitive equilibrium and planner's allocation will be different: $\left(\sigma_x - \sigma_h\right)^{CE} - \left(\sigma_x - \sigma_h\right)^{SP} = -\frac{1}{\gamma}\left(\sigma_\Lambda^{CE} - \sigma_\Lambda^{SP}\right)$. We can interpret this as an inefficient financial amplification channel, e.g., intermediaries are taking too much aggregate risk in the competitive equilibrium compared to the planner's allocation. In Section V I show this is in fact the case for uncertainty shocks that increase idiosyncratic risk ν_t. However, the analysis in this section shows that even in this case there is no inefficiency in the FOC for aggregate risk sharing. The planner can only improve the allocation by dealing with the externality in (33), and once this is done there is no need to further regulate intermediaries' risk taking decisions.

D. *Optimal Policy*

We can implement the optimal allocation as a competitive equilibrium with a tax on assets.[33] If an intermediary holds $q_t k_{i,t}$ in capital, he must pay a tax flow $\tau_t^k q_t k_t$ to the government.[34]

[31] This is the case, for example, in Brunnermeier and Sannikov (2014) and He and Krishnamurthy (2012, 2014a).

[32] This means that even if price of capital is technologically fixed, $\iota_t' = \bar{q}$, the competitive equilibrium with short-term contracts is still inefficient.

[33] Notice that while trading in capital markets cannot be controlled by the planner, intermediaries' asset holdings are observable and contractible.

[34] The policy instrument is not unique: we could also implement the optimal allocation with other instruments that reduce asset prices without distorting other private incentives.

PROPOSITION 1: *The planner's optimal allocation $\mathcal{P}$ can be implemented as part of a competitive equilibrium with a tax on asset holdings τ^k. To implement the optimal allocation, we need to set the present value of taxes relative to the market value of capital $T_t/q_t = \eta_t$.*

The tax τ^k reduces the equilibrium price of capital, as seen in the pricing equation (20). The planner is in fact using the tax τ^k to force agents to internalize the externality, so the planner's FOC (33) and the competitive equilibrium condition (35) align. This requires $T_t/q_t = \eta_t$, where $T_t k_t$ is the present value of taxes. Intuitively, the present value of future taxes reduces the equilibrium market value of capital q_t, internalizing the externality η_t. To balance the budget, the planner distributes the tax proceeds via lump-sum transfers, which are part of private agents' total wealth $(q_t + T_t)k_t$ (with complete markets this is equivalent to giving agents a government asset worth $T_t k_t$). The next section provides a sufficient statistic representation of the externality η_t in terms of equilibrium observables.

Policy Intervention versus Financial Amplification Channel.—It may seem surprising that the optimal policy intervention consists of taxing capital to reduce its equilibrium price, when lower asset prices are typically considered part of the financial amplification channel that creates financial crises. In fact, the setting here features precisely this type of financial amplification channel, as explained in Section IIB. The resolution to this puzzle lies in the endogenous response of intermediaries' net worth.

To fix ideas, consider the reduced-form equity constraint that arises from the moral hazard problem and restricts idiosyncratic risk sharing. Intermediaries are otherwise free to raise debt and share aggregate risk. Since intermediaries must retain at least a $\tilde{\phi}_t$ equity share, their exposure to idiosyncratic risk is

$$\tilde{\sigma}_{i,n,t} \geq \tilde{\phi}_t \frac{q_t k_{i,t}}{n_{i,t}} \nu_t. \tag{38}$$

The essence of the financial amplification channel in this environment is as follows. Suppose a bad aggregate shock hits the economy and the value of assets $q_t k_{i,t}$ falls; and suppose these financial losses are disproportionately concentrated on the balance sheets of intermediaries. As a result, intermediaries' net worth $n_{i,t}$ falls proportionally more than $q_t k_{i,t}$, and the constraint tightens. Intermediaries must be exposed to more idiosyncratic risk for the same capital, which drives up the required excess return on capital α_t (see equation (20)), further reducing asset prices q_t and tightening the constraint even more in a feedback loop.

In contrast, the social planner is taxing capital to reduce its equilibrium price q_t, but rebates the tax proceeds to agents. This allows him to reduce $q_t k_t$ without affecting n_t. This clearly relaxes the constraint: intermediaries are exposed to less idiosyncratic risk. The price of capital is still lower, so in both cases investment will fall. But here instead of tighter constraints that increase idiosyncratic risk, we get relaxed constraints and less idiosyncratic risk. It is this trade-off between investment and risk sharing that the externality η_t in equation (33) captures.

In many models, *aggregate* risk sharing is incomplete, so in addition to the externality studied here the MRS doesn't equalize across aggregate states.

The planner may therefore also wish to somewhat raise the price of capital in states of the world where the constraint is very tight, as a way of transferring wealth to intermediaries in those states (indirectly improving aggregate risk sharing).[35] But with complete aggregate risk sharing this is never necessary. If the planner wants to affect how aggregate risk is shared, he can simply regulate intermediaries' exposure to aggregate risk without distorting the price of capital. As it turns out, there is no reason to do this here because long-term contracts correctly evaluate the MRS Λ, as explained in Section IIIC.

The previous argument also shows how robust the intuition is. As long as intermediaries face an equity constraint that forces them to retain idiosyncratic risk as in (38), and the planner is able to use a tax or other such instrument to reduce the equilibrium price of capital q_t independently of intermediaries' net worth n_t, it will be advantageous to do so. This is true even if the equity stake $\tilde{\phi}_t$ is an endogenous object (in fact, this is the case in this paper).[36] Of course, if we just assume the reduced-form equity constraint we can't be sure that the policy intervention is consistent with the underlying contractual environment, or that it is optimal. The mechanism design approach in this paper allows us to see that it is indeed consistent with the underlying environment and optimal.

Tax on Assets versus Capital Requirements.—It is useful to distinguish the tax on asset holdings from capital requirements, which are a common component of financial regulation policy in practice. The tax on assets penalizes intermediaries for large asset holdings, but doesn't affect the equity versus debt margin. Intermediaries have private reasons to prefer debt or equity, and the planner doesn't need to interfere. Capital requirements, instead, impose minimum equity levels proportional to asset holdings: $(n_t + e_t) \geq \lambda_t q_t k_t$. They impose a penalty on issuing debt, but not equity, so they distort the debt/equity margin. To the extent that distorting this margin is costly, they also act as a penalty on assets, but only indirectly. In this environment, however, there is no reason to distort the debt/equity margin, so capital requirements are not optimal.

Capital requirements are often justified on the grounds that debt creates incentives for risk-shifting behavior by equity holders. However, outside equity holders don't have any moral hazard problem in this environment, so their incentives are not important. What matters is insiders' incentives, and private contracts correctly take them into account. The inefficiency in this economy does not arise from privately inefficient contracts, but rather from an externality that makes privately optimal contracts socially inefficient. Alternatively, if the government cannot commit to not bailout financial institutions such as banks, this creates incentives for excessive risk taking and may justify the use of capital requirements. This is the case in Chari and Kehoe (2013) for example. In this environment there is no need for bailouts, either ex ante or ex post.

[35] See Lorenzoni (2008) and Dávila and Korinek (2016).

[36] Here the contract can relax the equity stake $\tilde{\phi}_t$ by distorting the intertemporal consumption path, as discussed in Section IIA and equation (12) in particular. Reducing the price of capital q_t while keeping n_t constant improves risk sharing for a given equity stake $\tilde{\phi}_t$, or we can reduce the distortions in intertemporal consumption and still get the same idiosyncratic risk (or a little bit of both).

Time-Consistency of Optimal Policy.—We established in Section IIIB that the planner's allocation is renegotiation-proof in the sense that the only way of delivering more utility to households is to reduce intermediaries' utility, conditional on the exogenous aggregate state Y (in other words, $S'_X < 0$). This means optimal policy is time-consistent. The planner cannot obtain a Pareto improvement from deviating from his plan after any history. Of course, if he prefers households over intermediaries, he has incentives to expropriate them and transfer their wealth to households. But he can't deviate and make everyone better off.

IV. A Sufficient Statistic for the Externality and Optimal Policy

We can obtain a simple sufficient statistic for the externality in terms of equilibrium observables.

PROPOSITION 2: *For any competitive equilibrium*, *the externality satisfies the following expression:*

$$\eta_t = \alpha_t \epsilon_t, \tag{39}$$

where $\epsilon_t = \iota_t''(g_t)/\iota_t'(g_t)$ *is the semi-elasticity of the price of capital with respect to investment/growth.*

The formula (39) is true after any history, for any type of aggregate shocks, and it has a simple interpretation. The externality comes about from an un-internalized trade-off between investment/growth and idiosyncratic risk sharing. The planner realizes that if he wants more investment he must raise the price of capital, and this results in worse idiosyncratic risk sharing. The semi-elasticity ϵ_t captures the marginal effect of increasing g on the price of capital $\iota'(g)$, while α_t measures the marginal cost of exposing intermediaries to more idiosyncratic risk. Recall that α_t is the risk-adjusted expected excess return on capital, which compensates intermediaries for the idiosyncratic risk they must retain if they invest an extra dollar in capital. Taking them together we obtain the marginal cost of increasing g coming from idiosyncratic risk, in terms of reduced consumption for households. Expression (39) uses endogenous objects which may of course be affected by policy. However, it is always correctly measuring the size of the externality for any competitive allocation (for example, without taxes $\tau^k = 0$, or for the optimal taxes).

The model also provides an expression for the risk-adjusted expected excess return on capital α_t. From the pricing equation for capital (20) we obtain

$$\alpha_t = \gamma \frac{q_t k_t}{n_t} \left(\tilde{\phi}_t \nu_t\right)^2. \tag{40}$$

We can use this expression to understand how the externality (and optimal policy) responds to aggregate shocks. If we take ϵ as technologically fixed, the externality is larger after aggregate shocks that endogenously lead to weaker balance sheets for intermediaries (larger qk/n), tighter financial frictions $\tilde{\phi}$, or more idiosyncratic risk ν_t: or equivalently, aggregate shocks that raise the risk-adjusted expected excess

return α_t, e.g., downturns and financial crises. Optimal policy should therefore respond with higher taxes on assets after these aggregate shocks.

Do we really want to raise the tax on assets (and reduce their price q_t) precisely during a financial crisis? This is not as counterintuitive as it may sound at first. We use the tax to improve risk sharing, and we definitely want to improve risk sharing during a crisis. The main problem during a crisis is that risk is very high; low asset prices merely reflect this. Keep in mind that under the optimal policy capital is only a fraction of total wealth $(1 + T/q)qk$; while qk may go down with higher taxes, T/q goes up to compensate. In addition, a time-varying tax fixes incentives for the allocation of aggregate risk. By improving risk sharing and reducing excess returns during downturns, it corrects incentives for excessive exposure to aggregate risk ex ante. Intuitively, intermediaries are not that interested in taking large financial losses during downturns if their excess returns are going to be taxed away in that state. Raising the tax on assets during a financial crisis can help avoid the crisis in the first place.[37]

A. *Quantitative Analysis*

We can use market data to get a sense of the size of the externality. Since equation (39) is true after any history, and for any type of aggregate shocks, we can use it to compute the time varying externality if we have a time series for the risk adjusted expected excess return on assets α_t. Alternatively, equation (40) gives us an expression for α_t within the model, in terms of observables. Fix the semi-elasticity of the cost of capital $\epsilon = 3$ and a constant insiders' equity stake $\tilde{\phi} = 20$ percent, both from He and Krishnamurthy (2014). The calibration of insider's retained equity stake $\tilde{\phi} = 20$ percent is based on the common $2/20$ compensation scheme for hedge funds, PE, and VC funds; and from the fact that the average equity ownership of officers and directors in the finance, insurance, and real estate sectors is 17.4 percent.[38] Set the relative risk aversion $\gamma = 3$. We can extend the analysis in He, Kelly, and Manela (2017), to measure leverage in the financial sector at market values:[39]

$$\text{leverage}_t = \frac{\sum_i (\text{market equity}_{i,t} + \text{book debt}_{i,t})}{\sum_i \text{market equity}_{i,t}}.$$

I use CRSP/Compustat data with Standard Industrial Classification (SIC) codes 60–67, and use quarterly data from 1975:I to 2015:I. Notice that in the model $\text{leverage}_t = \frac{q_t k_t}{n_t + e_t} = \frac{q_t k_t}{n_t}\tilde{\phi}_t$. Figure 2 shows the resulting time series. The average leverage in the data is 10.8, and goes up during downturns. It was also relatively low between 1993 and 2006, and spiked during the financial crisis in 2008.

[37] But remember that the optimal allocation is renegotiation-proof. The planner is not going out of his way to punish intermediaries ex post to provide the right incentives ex ante: it is always fixing incentives looking forward.

[38] The retained equity stake is unlikely to be constant. In fact, the model predicts that it should become smaller during periods of financial distress. With more data we could in principle build a time series for insiders' equity stake $\tilde{\phi}_t$ and incorporate it into the analysis in a straightforward way. The structural approach in Section V addresses this issue.

[39] I thank Andres Schneider for the data and statistical analysis on leverage and idiosyncratic volatility.

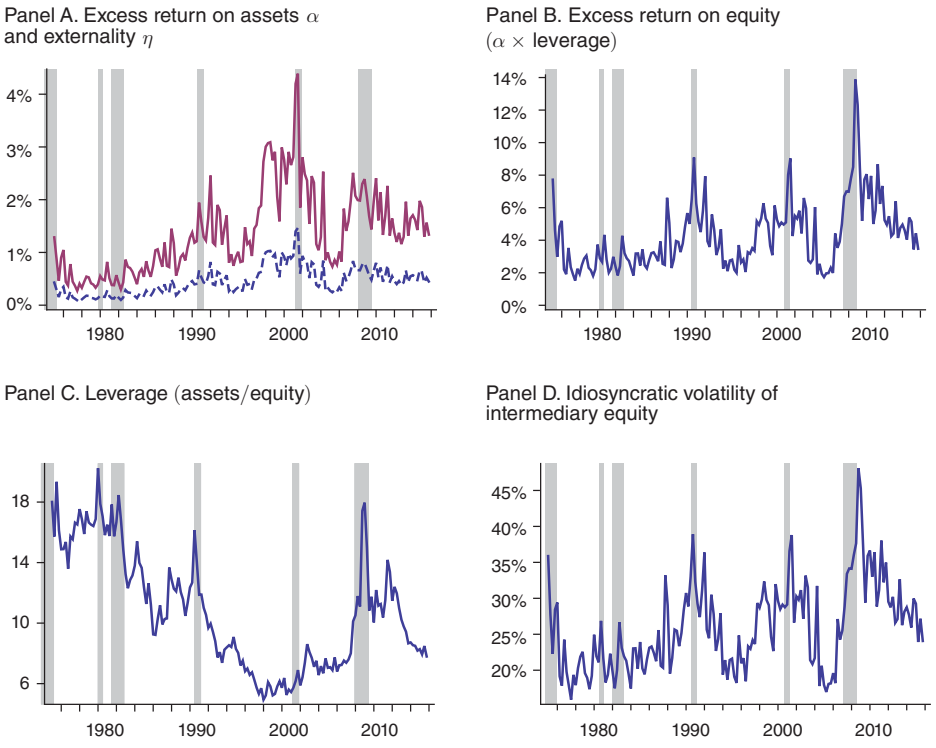

FIGURE 2

Notes: Panel A: risk-adjusted expected excess return on assets α (dashed) and externality η (solid). Panel B: risk-adjusted expected excess return on equity $\alpha \times$ leverage. Panel C: leverage. Panel D: idiosyncratic volatility of intermediary equity.

For the idiosyncratic risk of financial intermediaries I follow Herskovic et al. (2016). I run the regression

$$\Delta \ln(\text{market equity}_{i,t}) = \beta_0 + \beta_1 FF_{1,t} + \beta_2 FF_{2,t} + \beta_3 FF_{3,t} + \varepsilon_{i,t}$$

for every year, using monthly data (again for SIC codes 60–67), where FF_n are the Fama-French factors (market, high minus low (HML), small minus big (SMB)). Idiosyncratic volatility is then computed as the standard deviation of residuals for each quarter, and annualized to obtain a series $\nu_{equity,t}$ for the volatility of equity. Notice that in the model the volatility of total equity corresponds to $\nu_{equity,t} = \text{leverage}_t \times \nu_t$. Figure 2 shows the resulting time series. The average volatility in the data is 25.95 percent. It goes up during downturns, especially during the financial crisis in 2008.

I compute the risk-adjusted expected excess return on assets α predicted by the model, $\alpha_t = \gamma\tilde{\phi}\frac{\nu^2_{equity,t}}{\text{leverage}_t}$, and the size of the externality η_t.[40] Figure 2 shows the

[40] In Section VI I extend the framework to incorporate heterogeneous intermediaries and asset classes. Here I am assuming a common idiosyncratic volatility on equity ν_{equity}, so we can write

results. The average excess return on assets α_t is 0.45 percent, and the average externality η_t is 1.36 percent. They were particularly high during the dot-com boom and spiked during the crash, reaching 1.46 percent and 4.39 percent respectively. They then came down and spiked up again during the financial crises in 2007, reaching roughly 0.8 percent and 2.39 percent, and remained elevated since then. It is striking that, according to the model, the externality was higher during the dot-com crash than during the financial crisis. The reason is that although the idiosyncratic volatility on equity was higher during the crisis, so was the financial sectors' leverage. The unlevered idiosyncratic volatility on assets ν, which matters for the excess return α, was actually higher during the dot-com crash. I also compute the excess return on equity, leverage$_t \times \alpha_t$, with an average of 4.26 percent. It goes up during downturns and spikes during the financial crisis, reaching 13.88 percent in 2008. This is the result of not only a high excess return on assets α, but also very high leverage. Table 1 summarizes results and reports alternative specifications for high constant relative risk aversion (CRRA) γ, tight equity constraints $\tilde{\phi}$, and high elasticity of the cost of investment ϵ.

Is an average externality of 1.36 percent large or small? To put it in context, consider the required reduction in the investment share of GDP, I/Y, to reduce the marginal cost of capital by 1.36 percent. Assuming $K/Y = 3$ (Jones 2016) and $\epsilon = 3$, we can use Tobin's q to obtain a reduction in I/Y of 1.36 percentage points (e.g., from 20 percent to 18.64 percent of GDP). The externality computed here corresponds to an average across all asset classes. It is useful to get a sense of the overall quantitative importance of the mechanism. However, the externality is likely to vary significantly across asset classes. Some assets may require very high taxes, while others barely anything. Section VI extends the framework to heterogeneous intermediaries and asset classes.

B. *Implementation of the Optimal Allocation*

To implement the optimal allocation we must set $T_t/q_t = \eta_t$. Intuitively, the present value of future taxes T_t reduces the equilibrium market value of capital q_t. This makes investment less attractive and internalizes the externality, equating the planner's FOC (33) and the CE equilibrium condition (35). While Proposition 2 provides a simple expression for the externality in terms of observable variables, it does not specify exactly how the tax on assets τ_t^k or capital requirements should be set. In general, a time-varying tax τ_t^k will be required. We can recover the optimal tax τ_t^k from equation (16):

$$\tau_t^k = \eta_t \left(r_t - E_t^Q \left[\frac{d(\eta_t q_t k_t)}{\eta_t q_t k_t} \right] \right) \tag{41}$$

$$= \eta_t \big((r_t + \sigma_t \pi_t - g_t) - \mu_{\eta,t} - \mu_{q,t} - \sigma_{\eta,t}\sigma_{q,t} + (\pi_t - \sigma_t)(\sigma_{\eta,t} + \sigma_{q,t}) \big). \tag{42}$$

$\alpha = \sum_i \alpha_i \frac{q_t k_{i,t}}{q_t k_t} = \sum_i \gamma \tilde{\phi} \frac{\text{market equity}_{i,t}}{\text{market equity}_{i,t} + \text{book debt}_{i,t}} \nu^2_{equity,t} \frac{\text{market equity}_{i,t} + \text{book debt}_{i,t}}{\sum_i \text{market equity}_{i,t} + \text{book debt}_{i,t}} = \gamma \tilde{\phi} \frac{\nu^2_{equity,t}}{\text{leverage}_t}$.

Table 1—Average Values for the Baseline and Alternative Specifications

	Baseline	High γ	High $\tilde{\phi}$	High ϵ
CRRA γ	3	10	3	3
Equity stake $\tilde{\phi}$ (%)	20	20	50	20
Elasticity ϵ	3	3	3	6
Leverage $\frac{\text{assets}}{\text{total equity}}$	10.8			
Id. vol. eq. (%)	25.95			
Excess return α (%)	0.45	1.51	1.13	0.45
$\alpha \times$ leverage (%)	4.26	14.21	10.65	4.26
Externality η (%)	1.36	4.52	3.39	2.71

In a steady state $(\mu_\eta = \mu_q = \sigma_\eta = \sigma_q = 0)$ we would get a constant tax

$$\tau^k = \eta(r + \sigma\pi - g).$$

If we have a structural model that gives us the stochastic process for r, π, q, g, and α, we can use (41) to actually set the optimal tax τ_t^k. For example, if we calibrate $r + \sigma\pi = 7$ percent to match the average stock market return, $g = 2$ percent, and $\eta = 1.36$ percent, we obtain a tax $\tau^k \approx 7$ basis points.

In practice, however, it may be easier to follow a market-based approach. The planner can securitize the revenue from the tax on assets $\tau^k qk$, and let financial markets price it at $T_t k_t$. It can then measure T, q, and η, and adjust τ_t^k continuously to make sure $T_t/q_t = \eta_t$ as measured by (39).

V. A Numerical Example

In this section, I provide a numerical solution where the economy is hit by uncertainty shocks that increase idiosyncratic risk. The objective of this exercise is twofold. First, to illustrate the theoretical results with a concrete example. Second, to complement the quantitative evaluation of the externality in Section IV. An important advantage of the sufficient statistic approach is that we don't need to commit to specific parameter values or structural shocks. The results are valid for any type of aggregate shock and parameter values. But it also has drawbacks that a structural approach can overcome. First, the sufficient statistic is a local measure of the externality. If we actually introduce a tax on asset holdings to internalize it, the whole allocation will change. So we can't know what the value of the externality will be at the optimum. Second, if an input into the sufficient statistic can't be measured, a structural model can fill in the gap. This is in fact the case with the retained equity share $\tilde{\phi}_t$, which can be time-varying. Since we lack a good high-frequency measurement, in the quantitative evaluation in Section IV I used a constant $\tilde{\phi}_t = \tilde{\phi} = 20$ percent. But the theory suggests that optimal contracts should relax the retained equity share during downturns when idiosyncratic risk is high and

intermediaries' equity scarce. The structural model can help address these two issues, but comes at the cost of having to commit to a particular model. Each approach has advantages and drawbacks.

Aggregate Shocks and Parameter Values.—The economy is hit only by uncertainty shocks that increase idiosyncratic risk ν_t, which is the only exogenous state variable, $Y_t = \nu_t$, and follows an autoregressive process,

$$\underbrace{d\nu_t}_{d\widetilde{Y}_t} = \underbrace{\beta(\bar{\nu} - \nu_t)}_{\mu_Y(Y_t)}\, dt + \underbrace{\sqrt{\nu_t}\,\sigma_\nu}_{\sigma_Y(Y_t)}\, dZ_t.$$

As a convention, I will take $\sigma_\nu < 0$, so that we may think of Z as a "good" shock that drives idiosyncratic risk ν_t down.

I calibrate the model to make results comparable to the sufficient statistic exercise in Section IV. *Preferences*: I set the discount rate $\rho = 4$ percent and risk aversion $\gamma = 3$, which are standard, and EIS $\psi = 6$ as in He and Krishnamurthy (2014).[41] Intermediaries retire with Poisson intensity $\theta = 0.1$, so that their leverage in the steady state is roughly 10, as in the data in Section IV (this corresponds to a steady-state value of $X = 0.08$). *Moral hazard*: I set the intermediaries' efficiency of fund diversion $\phi = 0.39$ to that in the steady state the equilibrium retained equity share is $\tilde{\phi} = 20$ percent as in He and Krishnamurthy (2014), and in line with the common $2/20$ compensation scheme. *Technology*: I set the investment cost function $\iota(g) = (\exp(\epsilon g) - 1)/\epsilon + \delta$, with $\delta = 5$ percent and $\epsilon = 3$ as in He and Krishnamurthy (2014). I set the marginal product of capital $a = 0.1$ to obtain a growth rate in the steady state of 2 percent. I focus on uncertainty shocks so I set the volatility of TFP $\sigma = 0$. *Uncertainty shocks*: I set the long-run value $\bar{\nu} = 2.5$ percent so that the idiosyncratic risk of intermediary equity is roughly 25 percent in the steady state, as in the data in Section IV. I set the mean-reversion parameter $\beta = 0.065$, corresponding to a half-life of 10 years, and $\sigma_\nu = -0.1/\sqrt{10}$ so that the volatility of the idiosyncratic risk of intermediary equity, at the steady state, is $0.1\sqrt{\tilde{\sigma}_n}$.

Competitive Equilibrium versus Social Planner.—This numerical solution illustrates the theoretical results with a concrete example. Figure 3 shows that the social planner can deliver more utility to households S for any level of utility for intermediaries X. He achieves this by reducing asset prices q, which improves idiosyncratic risk sharing. This allows him to deliver utility to intermediaries at a lower cost, but requires lower investment and growth (investment is approximately $\iota = (q - 1)/\epsilon + \delta$). In both the unregulated competitive equilibrium and the planner's allocation, asset prices q (and therefore investment) are lower when experts' continuation utility X is low, and when idiosyncratic risk ν is high. It is

[41] A high EIS prevents interest rates from falling to stabilize the price of capital when the risk premium rises.

costly to provide incentives to intermediaries when capital is very risky relative to their continuation utility, so capital is less attractive.

Notice that we can always relax the retained equity share $\tilde{\phi}$ and improve idiosyncratic risk sharing by distorting the intertemporal consumption margin, and this is more attractive when X is low and ν high. But the planner already improves risk sharing by reducing the price of capital, so his incentives to distort intertemporal consumption to relax $\tilde{\phi}$ are weaker. As a result, $\tilde{\phi}$ is larger under the planner's allocation, and intermediaries' exposure to idiosyncratic risk $\tilde{\sigma}_n$ doesn't fall as much as it would have had we kept the retained equity share $\tilde{\phi}$ unchanged.

Figure 4 shows the externality η produced by hidden trade. It is larger when idiosyncratic risk ν is large and intermediaries' utility X is low. When intermediaries are highly exposed to idiosyncratic risk, the marginal cost of further increasing this exposure by raising investment and the price of capital is very large. The externality η can be computed both for the unregulated competitive equilibrium and the social planner's allocation, and the sufficient statistic (39), $\eta = \alpha\epsilon$, is valid in both allocations (in the planner's allocation it's internalized by the tax on assets). The externality η is a local concept that measures the wedge between the private and social FOC at any allocation. One limitation of the sufficient statistic approach used in Section IV is that if we measure η at the competitive equilibrium using the sufficient statistic, we cannot know how it will change once we actually introduce the optimal taxes; η gives only a local measure of the externality. But at least in this numerical solution, the externality η barely changes between the competitive equilibrium and the planner's solution. It is slightly higher in the planner's allocation. In the steady state, the externality η is 1.20 percent in the competitive equilibrium versus 1.21 percent in the planner's allocation.

Figure 4 also shows the MRS between intermediaries' and households' utility Λ. In the unregulated competitive equilibrium it is larger when intermediaries' utility X is higher and idiosyncratic risk ν lower. To understand this, recall from expression (25) for Λ that giving more consumption to intermediaries gives them more utility both directly, $\hat{c}_t^{-1/\psi}$, and indirectly by relaxing the IC constraints and improving idiosyncratic risk sharing, $\frac{\gamma}{\psi}\left(\phi_t \iota_t'(g_t)\frac{\nu_t}{X_t}\right)^2 \hat{c}_t^{-2/\psi-1}$. This second effect is small when X is large and ν low, because risk sharing is already relatively good. As a result, the MRS Λ is larger under these conditions.

This is also the reason why Λ is larger under the social planner's allocation. Since the planner already relaxes the IC constraint by distorting investment, the benefit of giving intermediaries more consumption is smaller, and the cost of giving them more utility is therefore larger. In fact, this also explains why the difference between the MRS Λ along the unregulated competitive equilibrium and the social planner's allocation is larger when idiosyncratic risk ν is high, a fact that will play an important role in the allocation of aggregate risk. Intuitively, when intermediaries are more exposed to idiosyncratic risk the planner's intervention is more potent. If idiosyncratic risk is very small, the planner's intervention barely matters, so the MRS Λ is similar. It is important to remember that we are dealing with "slopes," rather than "levels." Households do get more utility under the social planner's allocation as can be seen in Figure 3. But the *marginal* cost of giving more utility to intermediaries is larger.

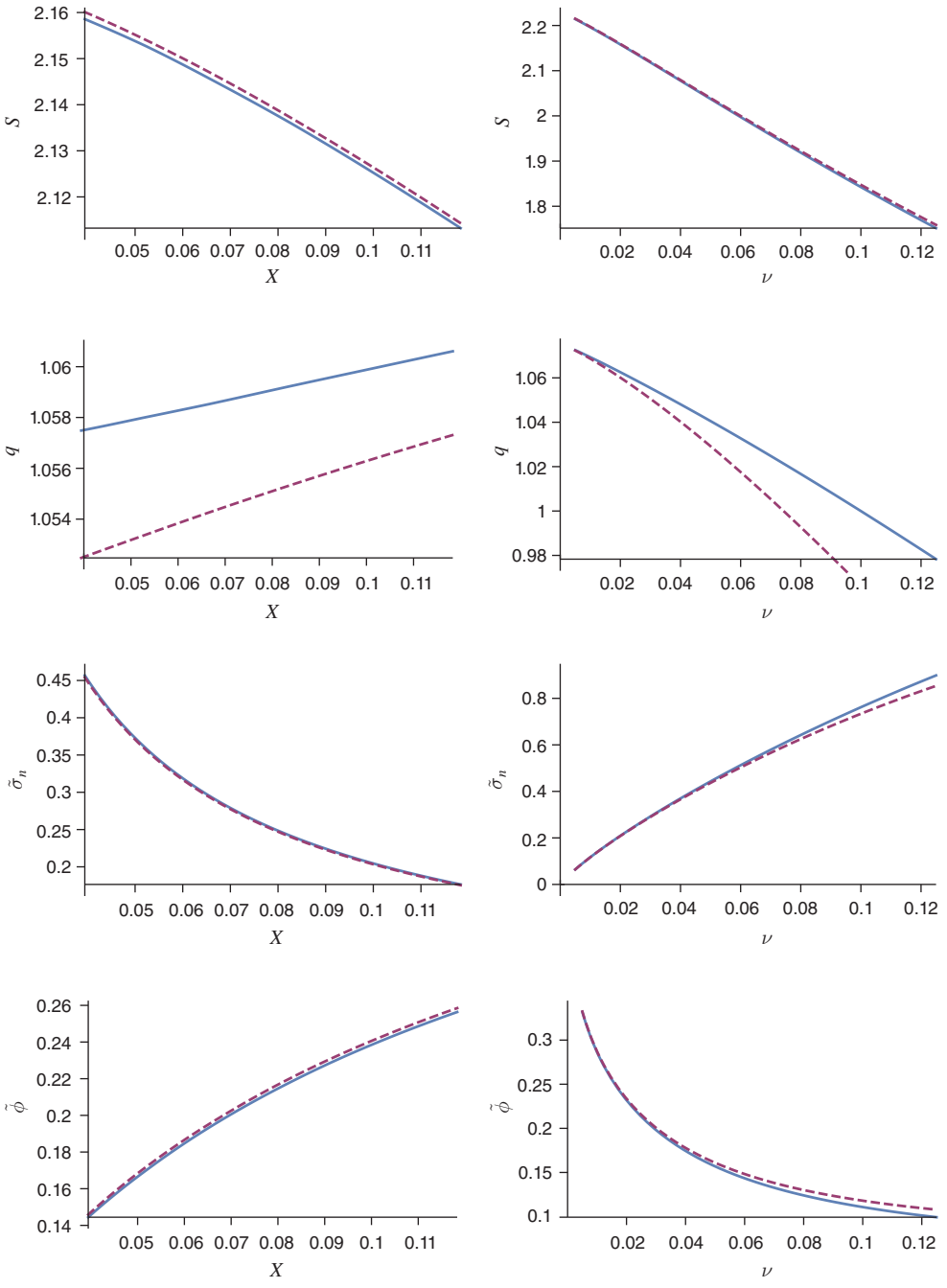

FIGURE 3

Notes: Households' utility S, price of capital q, intermediaries' idiosyncratic risk $\tilde{\sigma}_n$, and the retained equity share $\tilde{\phi}$, as function X for the steady state $\bar{\nu} = 2.5$ percent (left), and as functions of ν for the steady state $X = 0.08$ (right). Solid line is the CE, dashed line is the SP.

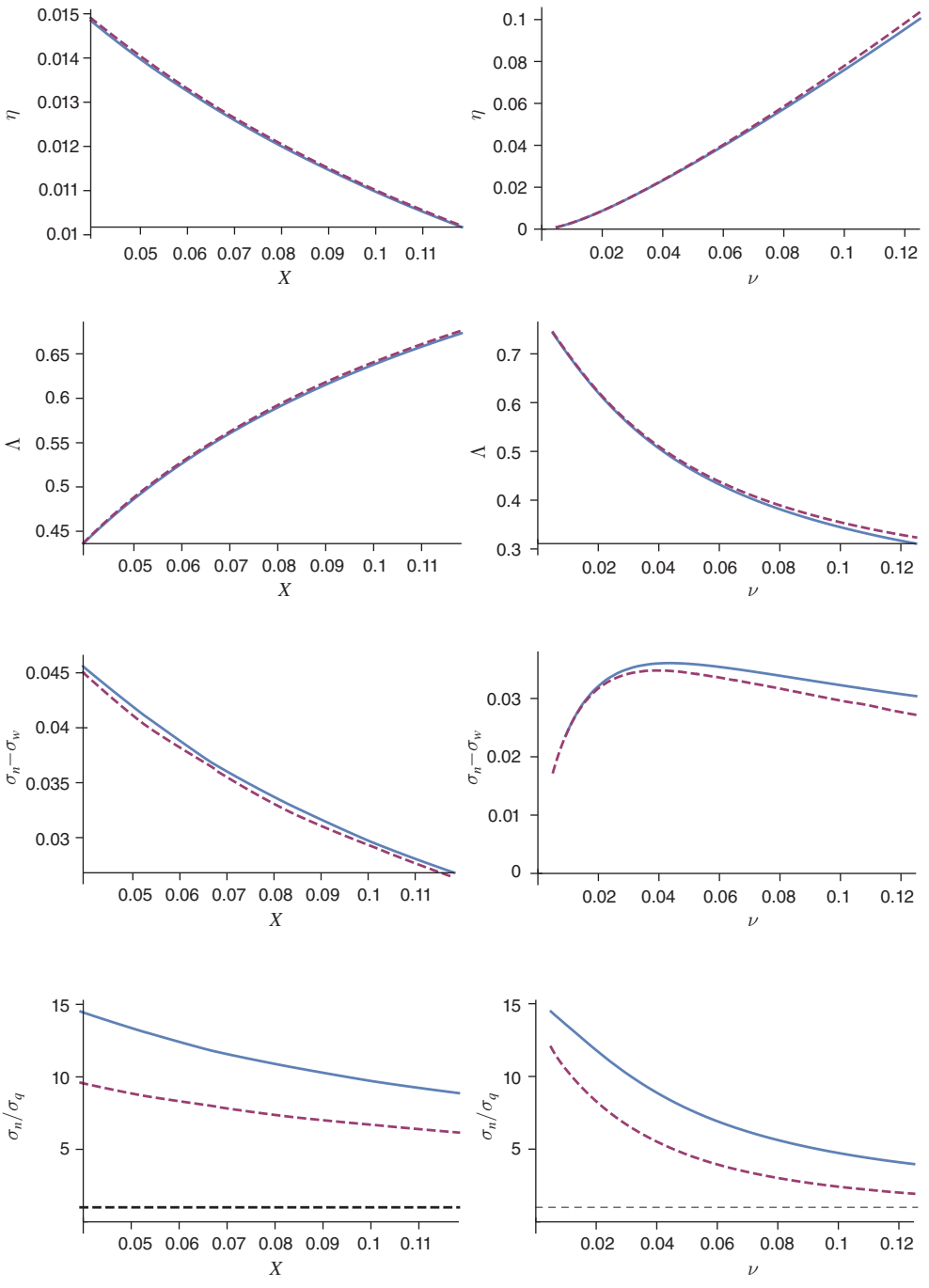

FIGURE 4

Notes: Externality η, MRS between intermediaries' and households utility Λ, sharing of aggregate financial risk $\sigma_n - \sigma_w$, and the amplification of risk on the balance sheets of intermediaries σ_n/σ_q, as functions of X for the steady state $\bar{\nu} = 2.5$ percent (left), and as functions of ν for the steady state $X = 0.08$ (right). Solid line is the CE, dashed line is the SP.

Figure 4 also shows how uncertainty shocks produce a financial amplification channel. After idiosyncratic risk ν goes up, the price of capital q falls and financial losses are concentrated on the balance sheets of intermediaries, $\sigma_n - \sigma_w > 0$. As explained in Section IIB, this drives the required excess return on capital α further up, depressing asset prices and investment even more. To understand why financial losses are concentrated on intermediaries, look at the behavior of the MRS Λ. With $\gamma > 1$, the FOC for aggregate risk sharing (21) says that financial losses are concentrated on intermediaries after aggregate shocks that reduce the cost of providing utility to them (shocks that reduce Λ). This happens after a bad uncertainty shock, as can be seen in Figure 4 and was explained above.

Although there is no externality associated with the FOC for aggregate risk sharing, this concentration of aggregate risk is excessive. In the social planner's allocation, financial losses are less concentrated on the balance sheets of intermediaries, $(\sigma_n - \sigma_w)^{CE} > (\sigma_n - \sigma_w)^{SP}$. The reason for this is that the marginal cost of delivering utility to intermediaries Λ doesn't go down as much after an uncertainty shock in the planner's allocation. As explained above, the planner responds to higher risk ν by lowering the price of capital q to relax the IC constraint and improve idiosyncratic risk sharing. Giving intermediaries more consumption still improves their risk sharing, but this is less valuable because idiosyncratic risk sharing is better than in the competitive equilibrium, so Λ falls less in the planner's allocation after idiosyncratic risk ν goes up.[42] As a result, the financial amplification channel is weaker in the planner's allocation. The last row of Figure 4 shows how risk is amplified on the balance sheets of intermediaries: if asset prices fall by x percent, intermediaries lose $(\sigma_n/\sigma_q) \times x$ percent of their net worth.

Quantitative Evaluation.—Switching to the optimal allocation yields on average[43] welfare gains equivalent to raising intermediaries' consumption permanently by 3.8 percent.[44] Asset prices are on average 0.38 percent lower in the planner's allocation, corresponding to roughly 0.13 percentage points lower growth. Intermediaries' idiosyncratic risk goes down on average by 0.59 percent when we switch to the planner's allocation. Aggregate risk is significantly concentrated on the balance sheets of intermediaries. In the competitive equilibrium, if asset prices go down by 1 percent, intermediaries lose 11 percent of their net worth on average. The optimal allocation reduces this concentration of aggregate risk. If asset prices fall by 1 percent intermediaries lose only 7.48 percent of their net worth on average, even though the planner is not intervening on this margin.

The average externality in the competitive equilibrium is 1.21 percent. As a reference, recall that the average externality we obtained using the sufficient statistic approach is 1.36 percent. The externality barely changes as we move from the

[42] This is reflected in the fact that $\tilde{\phi}$ is not only larger in the planner's allocation, but it also goes down less than in the competitive equilibrium when idiosyncratic risk ν rises.

[43] I use the stationary distribution of the competitive equilibrium to compare it to the planner's allocation. The stationary distribution of the planner's allocation is slightly different, but we are interested in an average pointwise comparison, rather than comparing averages. Figure 1 in the online Appendix shows the stationary distribution.

[44] This is considerably smaller than raising everyone's consumption by that amount, because intermediaries account for a small fraction of total consumption.

unregulated competitive equilibrium to the optimal allocation; it's also 1.21 percent on average. This helps us address one of the drawbacks of the sufficient statistic approach; namely, the fact that it is a local measure that is likely to change as we move to the optimal allocation. While the effect is there, it is negligible, at least for this structural model.

The average retained equity stake in the competitive equilibrium is 21.51 percent; it goes slightly up in the planner's allocation, to 21.72 percent, as explained above. In the sufficient statistic approach I used a constant retained equity stake of 20 percent, in line with the evidence in He and Krishnamurthy (2014), for lack of more detailed high-frequency data. But the model predicts that optimal contracts will relax the retained equity share after uncertainty shocks that raise idiosyncratic risk, as shown in Figure 3. This effect dampens the impact of uncertainty shocks, but the effect on the average externality is small. If we measured η on the model-generated data assuming that the retained equity stake $\tilde{\phi}$ is constant and equal to the steady-state value, I would obtain an average η of 1.23 percent, slightly above the actual average of 1.21 percent.

We can also look at what happens on impact after a large (and rare) uncertainty shock that triples idiosyncratic risk ν_t from its steady-state value 2.5 percent to 7.5 percent. Asset prices fall by only 3.6 percent, but intermediaries' net worth falls by 29.2 percent. Intermediaries' idiosyncratic risk goes up from 25.2 percent to 60 percent, and the retained equity share $\tilde{\phi}$ falls from 21.3 percent to 13 percent. The measured externality η goes from 1.20 percent to 5.29 percent. As a reference, the externality η measured in the sufficient statistic approach in Section IV reaches a maximum of 4.39 percent at the height of the dot-com boom.

Under the planner's allocation, instead, the same shock causes asset prices to fall by 5.4 percent, but intermediaries' net worth falls by 26 percent, less than in the competitive equilibrium because aggregate risk is less concentrated on the balance sheets of intermediaries. Intermediaries' idiosyncratic risk goes from 25 percent to 58 percent; their retained equity stake goes from 21.5 percent to 13.6 percent; and the externality η goes from 1.21 percent to 5.39 percent. Overall, the values for the competitive equilibrium and the planner are close, but the difference widens after an uncertainty shock. This is when the planner's intervention is more important.

VI. Heterogeneous Assets and Intermediaries

A practical concern for regulators is how to treat different asset classes (e.g., appropriate risk weighting), and different intermediaries. To address this issue, we can extend the model to incorporate heterogeneous asset classes and intermediaries. The main conclusions in the baseline model go through. The sufficient statistic formula (39) is valid for each asset class, and can be used to determine regulation policies at a more desegregated level. I develop the setting in detail in the online Appendix.

Suppose there are $F \geq 1$ types of intermediaries, and $J \geq 1$ asset classes. Assets may differ in their investment technology $\iota_{j,t}(g_{j,t})$ or their exposure to aggregate risk $\sigma_{j,t}$. Each type of intermediary may be able to deal with each asset

class differently, i.e., they get different idiosyncratic risk $\nu_{i,j,t}$, moral hazard $\phi_{i,j,t}$, or output flow $a_{i,j,t}$. Intermediary i gets observable return $R_{i,j,t}$ per dollar invested in asset j:

$$dR_{i,j,t} = \left(\frac{a_{i,j,t} - \iota_{j,t}(g_{j,t})}{q_{j,t}} + g_{j,t} + \mu_{q,j,t} + \sigma_{j,t}\sigma'_{q,j,t} - \tau^k_{j,t} - s_{i,j,t}\right)dt$$
$$+ (\sigma_{j,t} + \sigma_{q,j,t})\,dZ_t + \nu_{i,j,t}dW_{i,j,t},$$

where $W_{i,j}$ is an intermediary and asset class-specific idiosyncratic shock. Because the intermediary can secretly divert returns for asset class j, he must keep an exposure $\tilde{\sigma}_{n,i,j,t} = \tilde{\phi}_t \nu_{i,j,t}\frac{q_{j,t}k_{i,j,t}}{n_{i,t}}$ to $W_{i,j}$, with $\tilde{\phi}_{i,j,t} = \xi_{i,t}\hat{c}_{i,t}^{-1/\psi}\phi_{i,j,t}$.[45] As a result, we obtain the following asset pricing equation for each asset class j:

$$\underbrace{\frac{a_{i,j,t} - \iota_{j,t}(g_{j,t})}{q_{j,t}} + g_{j,t} + \mu_{q,j,t} + \sigma_{j,t}\sigma'_{q,j,t} - \left(r_t + \tau^k_{j,t}\right) - \left(\sigma_{j,t} + \sigma_{q,j,t}\right)\pi_t}_{\text{risk-adjusted excess return } \equiv\ \alpha_{i,j,t}}$$

$$= \underbrace{\gamma\frac{q_{j,t}k_{i,j,t}}{n_{i,t}}\left(\tilde{\phi}_{i,j,t}\nu_{i,j,t}\right)^2}_{\text{id. risk premium}}$$

for all (i,j) pairs such that $k_{i,j,t} > 0$. Intermediaries will invest more heavily in assets for which they are better suited (low $\phi_{i,j,t}$ and $\nu_{i,j,t}$, or high $a_{i,j,t}$). Reorganizing, we get $a_{i,j,t} - \gamma\frac{q_{j,t}k_{i,j,t}}{n_{i,t}}\left(\tilde{\phi}_{i,j,t}\nu_{i,j,t}\right)^2 q_{j,t}$ constant for all intermediaries who hold the asset. The market allocates assets to equalize the marginal benefit from output net of the cost of idiosyncratic risk. This is exactly what the planner would do, so there is no inefficiency in the allocation of assets across intermediaries. This also means that we can implement the optimal allocation with asset specific taxes $\tau^k_{j,t}$ which treat all intermediaries the same.

But how should different asset classes be treated? The externality now takes into account how increasing the marginal cost of capital of class j, $\iota'_{j,t}(g_j)$, tightens the idiosyncratic risk sharing problem of every type of intermediary:

$$\eta_{j,t} = \alpha_{j,t}\epsilon_{j,t}, \tag{43}$$

where $\alpha_{j,t} = \sum_i \alpha_{i,j,t}\frac{q_{j,t}k_{i,j,t}}{q_{j,t}k_{j,t}}$ is the value-weighted average risk-adjusted expected excess return on asset class j across all intermediaries, and $\epsilon_{j,t} = \frac{\iota''_{j,t}(g_{j,t})}{\iota'_{j,t}(g_{j,t})}$ the

[45] Notice that because the asset specific risks $\{W_{i,j}\}$ are independent there is some diversification. The total idiosyncratic volatility for intermediary i is $\sqrt{\sum_j\left(\tilde{\phi}_{i,j,t}\nu_{i,j,t}\frac{q_{j,t}k_{i,j,t}}{\sum_j q_{j,t}k_{i,j,t}}\right)^2}\left(\frac{\sum_j q_{j,t}k_{i,j,t}}{n_{i,t}}\right)$, so that even if $\tilde{\phi}_{i,j,t} = \tilde{\phi}_{i,t}$ and $\nu_{i,j,t} = \nu_{i,t}$, we get less idiosyncratic volatility than $\tilde{\phi}_{i,t}\nu_{i,t}\left(\frac{\sum_j q_{j,t}k_{i,j,t}}{n_{i,t}}\right)$.

semi-elasticity of $q_{j,t}$ with respect to $g_{j,t}$. The excess return $\alpha_{j,t}$ measures the cost, from idiosyncratic risk, of raising the value of assets of class j, while $\epsilon_{j,t}$ measures how much we must raise asset values to increase the growth rate $g_{j,t}$. Expression (43) measures the externality for any competitive equilibrium, even if we are not implementing the planner's optimal allocation. The optimal policy must set the value of the tax on each asset class equal to the externality in that class, $T_{j,t}/q_{j,t} = \eta_{j,t}$.

PROPOSITION 3: *With heterogeneous assets classes and intermediaries, the planner's optimal allocation $\mathcal{P}$ can be implemented with an asset-specific tax on asset holdings $\left\{\tau_j^k\right\}$ that treats all intermediaries the same. The optimal tax internalizes the externality $T_{j,t}/q_{j,t} = \eta_{j,t}$ satisfying (43) in equilibrium.*

Equation (43) tells us how regulators should treat different asset classes. Regulators don't need to concern themselves with the riskiness of each asset class, or even their systemic risk. Instead, the average excess return $\alpha_{j,t}$ contains all the relevant information, and reflects both the riskiness of the asset class and the place it occupies on intermediaries' balance sheets.

We can use an appropriately modified version of equation (41) to set $\tau_{j,t}^k$, and use the same market based implementation in Section IV. To measure the externality on all capital, analogous to expression (39) in the previous section, we need to take into account the correlation between the excess return $\alpha_{j,t}$ and the semi-elasticity $\epsilon_{j,t}$:

$$\frac{\sum_j \eta_{j,t} q_{j,t} k_{j,t}}{\sum_j q_{j,t} k_{j,t}} = \alpha_t \epsilon_t + \sum_j \big(\alpha_{j,t} - \alpha_t\big)\big(\epsilon_{j,t} - \epsilon_t\big) \frac{q_{j,t} k_{j,t}}{\sum_j q_{j,t} k_{j,t}},$$

where $\alpha_t = \sum_j \alpha_{j,t} \frac{q_{j,t} k_{j,t}}{\sum_j q_{j,t} k_{j,t}} = \sum_i \alpha_{i,t} \frac{\sum_j q_{j,t} k_{i,j,t}}{\sum_j q_{j,t} k_{j,t}}$ is the value-weighted excess return on all capital (which can be measured across asset classes or across intermediaries), and $\epsilon_t = \sum_j \epsilon_{j,t} \frac{q_{j,t} k_{j,t}}{\sum_j q_{j,t} k_{j,t}}$ is the value-weighted semi-elasticity of the price of capital with respect to growth. The first term on the right-hand side corresponds to the one in equation (39). The second term captures the value-weighted covariance between excess return $\alpha_{j,t}$ and semi-elasticity $\epsilon_{j,t}$. In the special, but salient, case where $\alpha_{j,t}$ and $\epsilon_{j,t}$ are uncorrelated, we recover expression (39).

VII. Conclusion

In this paper, I propose a model of optimal financial regulation where financial intermediaries trade capital assets on behalf of households, but must retain an equity stake for incentive reasons. This is a commonly observed financial arrangement, and widely used in models of financial crises. Financial regulation is necessary in this environment because intermediaries cannot be excluded from privately trading in capital markets. Private intermediaries don't internalize that when they demand assets and bid up their price they force others to bear more idiosyncratic risk.

The main takeaway is that the optimal allocation requires lower asset prices in order to improve risk sharing, even if it comes at the cost of lower investment. Essentially, low asset prices make the whole financial system less risky. The optimal allocation can be implemented with a tax on asset holdings that internalizes the hidden-trade externality. Once this is done, there is no need for further regulation. The externality admits a simple sufficient statistic representation that allows us to measure it using market data. I find the externality represents 1.36 percent of the market value of assets on average, but it spikes during downturns and financial crises, where it can reach 4.39 percent. While these are average values across all assets, the formula is valid for each asset class, and can be used to determine how different asset classes should be treated.

The competitive equilibrium may feature an inefficient financial amplification channel, in the sense that intermediaries may have an excessive exposure to aggregate risk compared to the socially optimal allocation. However, there is no need to directly regulate intermediaries' exposure to aggregate risk. Once the hidden trade externality is internalized, privately optimal contracts correctly allocate aggregate risk. Likewise, while the planner wants to tax intermediaries for their asset holdings, it doesn't want to distort the equity/debt margin (e.g., doesn't want to use capital requirements).

REFERENCES

Ackermann, Carl, Richard McEnally, and David Ravencraft. 1999. "The Performance of Hedge Funds: Risk, Return, and Incentives." *Journal of Finance* 54 (3): 833–74.

Alvarez, Fernando, and Urban J. Jermann. 2000. "Efficiency, Equilibrium, and Asset Pricing with Risk of Default." *Econometrica* 68 (4): 775-97.

Ang, Andrew. 2014. *Asset Management: A Systematic Approach to Factor Investing*. Oxford: Oxford University Press.

Bansal, Ravi, Dana Kiku, Ivan Shaliastovich, and Amir Yaron. 2014. "Volatility, the Macroeconomy, and Asset Prices." *Journal of Finance* 69 (6): 2471–511.

Beeler, Jason, and John Y. Campbell. 2009. "The Long-Run Risks Model and Aggregate Asset Prices: An Empirical Assessment." NBER Working Paper 14788.

Berk, Jonathan B., and Richard C. Green. 2004. "Mutual Fund Flows and Performance in Rational Markets." *Journal of Political Economy* 112 (6): 1269–95.

Bernanke, Ben, Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." In *Handbook of Macroeconomics*, Vol. 1, edited by John Taylor and Michael Wood, 1341–93. Amsterdam: Elsevier.

Biais, Bruno, Thomas Mariotti, Guillaume Plantin, and Jean-Charles Rochet. 2007. "Dynamic Security Design: Convergence to Continuous Time and Asset Pricing Implications." *Review of Economic Studies* 74 (2): 345–90.

Bianchi, Javier. 2011. "Overborrowing and Systemic Externalities in the Business Cycle." *American Economic Review* 101 (7): 3400–26.

Bianchi, Javier, and Enrique G. Mendoza. 2011. "Overborrowing, Financial Crisis, and 'Macro-Prudential' Policy." International Monetary Fund Working Paper 11/24.

Brunnermeier, Markus K., and Yuliy Sannikov. 2014. "A Macroeconomic Model with a Financial Sector." *American Economic Review* 104 (2): 379–421.

Chari, Varadarajan V., and Patrick J. Kehoe. 2013. "Bailouts, Time Inconsistency, and Optimal Regulation." NBER Working Paper 19192.

Chen, Qi, Itay Goldstein, and Wei Jiang. 2008. "Directors' Ownership in the U.S. Mutual Fund Industry." *Journal of Finance* 63 (3): 2629–677.

Dávila, Julio, Jay H. Hong, Per Krusell, and José-Víctor Ríos-Rull. 2012. "Constrained Efficiency in the Neoclassical Growth Model with Uninsurable Idiosyncratic Shocks." *Econometrica* 80 (6): 2431–67.

Dávila, Eduardo, and Anton Korinek. 2016. "Fire-Sales Externalities." NBER Working Paper 22444.

DeMarzo, Peter M., and Michael J. Fishman. 2007. "Optimal Long-Term Financial Contracting." *Review of Financial Studies* 20 (6): 2079–128.

DeMarzo, Peter M., Michael J. Fishman, Zhiguo He, and Neng Wang. 2012. "Dynamic Agency and the q Theory of Investment." *Journal of Finance* 67 (6): 2295–340.

DeMarzo, Peter M., and Yuliy Sannikov. 2006. "Optimal Security Design and Dynamic Capital Structure in a Continuous-Time Agency Model." *Journal of Finance* 61 (6): 2681–724.

Di Tella, Sebastian. 2017. "Uncertainty Shocks and Balance Sheet Recessions." *Journal of Political Economy* 125 (6): 2038–81.

Di Tella, Sebastian. 2019. "Optimal Regulation of Financial Intermediaries: Dataset." *American Economic Review.* https://doi.org/10.1257/aer.20161488.

Di Tella, Sebastian, and Yuliy Sannikov. 2016. "Optimal Asset Management Contracts with Hidden Savings." Unpublished.

Farhi, Emmanuel, Mikhail Golosov, and Aleh Tsyvinski. 2009. "A Theory of Liquidity and Regulation of Financial Intermediation." *Review of Economic Studies* 76 (3): 973–92.

Farhi, Emmanuel, and Iván Werning. 2016. "A Theory of Macroprudential Policies in the Presence of Nominal Rigidities." *Econometrica* 84 (5): 1645–704.

Gaspar, José-Miguel, Massimo Massa, and Pedro Matos. 2006. "Favoritism in Mutual Fund Families? Evidence on Strategic Cross-Fund Subsidization." *Journal of Finance* 61 (1): 73–104.

Geanakoplos, John D., Michael Magill, Martine Quinzii, and Jacques Dreze. 1990. "Generic Inefficiency of Stock Market Equilibrium When Markets Are Incomplete." *Journal of Mathematical Economics* 19 (1–2): 113–51.

Geanakoplos John D., and Heracles M. Polemarchakis. 1986. "Existence, Regularity, and Constrained Suboptimality of Competitive Allocations When the Asset Market Is Incomplete." In *Uncertainty, Information, and Communication: Essays in Honor of Kenneth J. Arrow*, Vol. III, edited by W. Heller, R. Starr, and D. Starrett, 65–95. Cambridge, UK: Cambridge University Press.

Gersbach, Hans, and Jean-Charles Rochet. 2012. "Aggregate Investment Externalities and Macroprudential Regulation." *Journal of Money, Credit, and Banking* 44: 73–109.

Gruber, Jonathon. 2013. "A Tax-Based Estimate of the Elasticity of Intertemporal Substitution." *Quarterly Journal of Finance* 31 (1): 1–20.

Hall, Robert E. 1988. "Intertemporal Substitution in Consumption." *Journal of Political Economy* 96 (2): 339–57.

Hart, Oliver D. 1975. "On the Optimality of Equilibrium When the Market Structure Is Incomplete." *Journal of Economic Theory* 11 (3): 418–43.

He, Zhiguo. 2012. "Dynamic Compensation Contracts with Private Savings." *Review of Financial Studies* 25 (5): 1494–549.

He, Zhiguo, Bryan T. Kelly, and Asaf Manela. 2017. "Intermediary Asset Pricing: New Evidence from Many Asset Classes." *Journal of Financial Economics* 126 (1): 1–35.

He, Zhiguo, and Peter Kondor. 2016. "Inefficient Investment Waves." *Econometrica* 84 (2): 735–80.

He, Zhiguo, and Arvind Krishnamurthy. 2012. "A Model of Capital and Crises." *Review of Economic Studies* 79 (2): 735–77.

He, Zhiguo, and Arvind Krishnamurthy. 2013. "Intermediary Asset Pricing." *American Economic Review* 103 (2): 732–70.

He, Zhiguo, and Arvind Krishnamurthy. 2014. "A Macroeconomic Framework for Quantifying Systemic Risk." NBER Working Paper 19885.

Herskovic, Bernard, Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh. 2016. "The Common Factor in Idiosyncratic Volatility: Quantitative Asset Pricing Implications." *Journal of Financial Economics* 119 (2): 249–83.

Jones, Charles I. 2016. "The Facts of Economic Growth." In *Handbook of Macroeconomics*, Vol. 2, 3–69. Amsterdam: Elsevier.

Kehoe, Timothy J., and David K. Levine. 1993. "Debt-Constrained Asset Markets." *Review of Economic Studies* 60 (4): 865–88.

Kiyotaki, Nobuhiro, and John Moore. 1997. "Credit Cycles." *Journal of Political Economy* 105 (2): 211–48.

Korinek, Anton. 2012. "Systematic Risk-Taking: Amplification Effects, Externalities, and Regulation Responses." Unpublished.

Lack, Simon. 2012. "The Hedge Fund Mirage: The Illusion of Big Money and Why It's Too Good to Be True." *CFA Institute Conference Proceedings Quarterly* 29 (4): 14–23.

Lorenzoni, Guido. 2008. "Inefficient Credit Booms." *Review of Economic Studies* 75 (3): 809–33.

Ma, Linlin, Yuehua Tang, and Juan-Pedro Gómez. 2015. "Portfolio Manger Compensation in the US Mutual Fund Industry." Paper presented at Finance Down Under Conference, Melbourne, Australia.

Mulligan, Casey B. 2002. "Capital, Interest, and Aggregate Intertemporal Substitution." NBER Working Paper 9373.

Phalippou, Ludovic. 2009. "Symposium: Private Equity: Beware of Venturing into Private Equity." *Journal of Economic Perspectives* 23 (1): 147–66.

Rampini, Adriano, and S. Viswanathan. 2010. "Collateral, Risk Management, and the Distribution of Debt Capacity." *Journal of Finance* 65 (6): 2293–322.

Sannikov, Yuliy. 2008. "A Continuous-Time Version of the Principal-Agent Problem." *Review of Economic Studies* 75 (3): 957–84.

Stiglitz, Joseph E. 1982. "The Inefficiency of the Stock Market Equilibrium." *Review of Economic Studies* 49 (2): 241–61.

Sun, Yeneng. 2006. "The Exact Law of Large Numbers via Fubini Extension and Characterization of Insurable Risks." *Journal of Economic Theory* 126 (1): 31–69.

Vissing-Jogensen, Annette. 2002. "Limited Asset Market Participation and the Elasticity of Intertemporal Substitution." *Journal of Political Economy* 110 (4): 825–53.

Zitzewitz, Eric. 2003. "How Widespread Is Late Trading in Mutual Funds?" Standford GSB Research Paper 1817.

This article has been cited by:

1. Felipe S. Iachan, Dejanir Silva, Chao Zi. 2021. Under-diversification and idiosyncratic risk externalities. *Journal of Financial Economics* **80**. . [Crossref]
2. Patrick Fève, Pablo Garcia Sanchez, Alban Moura, Olivier Pierrard. 2021. Costly default and skewed business cycles. *European Economic Review* **132**, 103630. [Crossref]
3. Alejandro Van der Ghote. 2021. Interactions and Coordination between Monetary and Macroprudential Policies. *American Economic Journal: Macroeconomics* **13**:1, 1-34. [Abstract] [View PDF article] [PDF with links]
4. Sebastian Di Tella, Yuliy Sannikov. 2021. Optimal Asset Management Contracts With Hidden Savings. *Econometrica* **89**:3, 1099-1139. [Crossref]
5. Joseph Lee, Yonghui Bao. Conflict of Goals in Takeover Law: The Impossible Regulatory Alignment Between UK and China 15-51. [Crossref]
6. Sebastian Di Tella. 2020. Risk Premia and the Real Effects of Money. *American Economic Review* **110**:7, 1995-2040. [Abstract] [View PDF article] [PDF with links]
7. Gregor Schwerhoff, Ottmar Edenhofer, Marc Fleurbaey. 2020. TAXATION OF ECONOMIC RENTS. *Journal of Economic Surveys* **34**:2, 398-423. [Crossref]
8. Nicholas Z. Muller. 2020. Long-Run Environmental Accounting in the US Economy. *Environmental and Energy Policy and the Economy* **1**, 158-191. [Crossref]
9. Gianluca Benigno, Huigang Chen, Christopher Otrok, Alessandro Rebucci, Eric R. Young. 2019. Optimal Policy for Macro-Financial Stability. *SSRN Electronic Journal* . [Crossref]
10. Stefano Pegoraro. 2019. Flows and Performance with Optimal Money Management Contracts. *SSRN Electronic Journal* . [Crossref]
11. Wenhao Li. 2018. Public Liquidity Supply, Bank Run Risks, and Financial Crises. *SSRN Electronic Journal* . [Crossref]