Working Paper Series

Asset Issuance in Over-The-Counter Markets

WP 17-13

Zachary Bethune
University of Virginia

Bruno Sultanum
Federal Reserve Bank of Richmond

Nicholas Trachter
Federal Reserve Bank of Richmond

FEDERAL RESERVE BANK OF RICHMOND®
Richmond ▪ Baltimore ▪ Charlotte

Asset Issuance in Over-The-Counter Markets*

Zachary Bethune
University of Virginia

Bruno Sultanum
Federal Reserve Bank of Richmond

Nicholas Trachter
Federal Reserve Bank of Richmond

October 19, 2017
Working Paper No. 17-13

Abstract

We model asset issuance in over-the-counter markets. Investors buy newly issued assets in a primary market and trade existing assets in a secondary market, where both markets are over the counter. We show that the level of asset issuance and its efficiency depend on how investors split the surplus in secondary market trade. If buyers get most of the surplus in secondary market trade, then sellers do not have incentives to participate in the primary market in order to intermediate assets and the economy has a low level of assets. On the other hand, if sellers get most of the surplus, buyers have strong incentives to participate in the primary market and the economy has a high level of assets. Equilibrium is inefficient for any splitting rule. The result follows from a double-sided hold-up problem in which it is impossible for all investors to take into account the full social value of an asset when trading. We propose a tax/subsidy scheme and show how it restores efficiency. We calibrate our model to match features of the US municipal bond market in order to quantify the effects of the intervention. The intervention leads to large welfare gains and, in response to a financial crisis caused by an aggregate demand shock, makes the crisis less severe and shorter relative to the economy with no intervention.

JEL Classification: D53, D82, G14

Keywords: Decentralized markets, bilateral trade, asset issuance, liquidity.

*Contact: Zachary Bethune: zab2t@virginia.edu, Bruno Sultanum: bruno@sultanum.com, Nicholas Trachter: trachter@gmail.com. Any opinions or views expressed in this article are not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

DOI: https://doi.org/10.21144/wp17-13

1 Introduction

Many assets, both real and financial, are traded in secondary over-the-counter (OTC) markets after their initial issuance (e.g. real estate, municipal bonds, treasuries, asset-backed securities, etc.). Further, many of these markets experienced severe volatility during the 2008 financial crisis, and several policies were enacted that aimed to directly support the issuance of new assets. For example, the Federal Reserve created the Term Asset-Backed Securities Loan Facility (TALF) to support the issuance of asset-backed securities collateralized by different types of private loans, and the Commercial Paper Funding Facility (CPFF) to support the issuance of commercial papers.[1] While there is a large literature studying OTC markets (see Duffie et al. (2005), Duffie et al. (2007), Lagos and Rocheteau (2009a), and Hugonnier et al. (2014), among other), most studies in this literature assume a fixed supply of assets, a nonstarter in understanding the effects of policies aimed to spur issuance. In this paper, we study how the trading of seasoned assets in secondary OTC markets affects their primary issuance and, in turn, aggregate asset supply and welfare.

We emphasize two frictions in OTC trade: (i) searching for counterparties to trade with takes time and (ii) conditional on a trade opportunity, the terms of the trade are determined by bargaining. We explore the canonical economy of Duffie et al. (2005) but for two differences: (i) we abstract from competitive market-makers in order to make the model more tractable, and (ii) we introduce the notion of issuers—agents who have a technology to issue new assets. In the model, trade occurs in pairwise meetings and these meetings are subject to frictions. We interpret meetings between an investor and an issuer as occurring in the primary market since they involve the potential issuance of a new asset. Likewise, we interpret meetings between two investors as occurring in the secondary market as they involve a transfer of a previously issued asset. When two agents meet, either in the primary or secondary market, the terms of trade are determined by Nash bargaining.

We solve for the decentralized equilibrium in the economy and compare it with the constrained efficient allocation, which is the welfare maximizing allocation constrained by the search frictions. Surprisingly, we find that the decentralized equilibrium allocation is never constrained efficient. This conclusion holds even though trade in the secondary market is constrained efficient when the asset level is fixed, consistent with the literature cited above. Under search and bargaining, the prices investors trade in the secondary market do not reflect the social return of assets. When asset supply is fixed, this mispricing is irrelevant in determining the allocation—assets flow from low-valuation agents to high-valuation agents—and equilibrium is constrained efficient. When we introduce issuance,

[1] See www.newyorkfed.org/markets/funding_archive/index.html for details.

this mispricing affects investors' incentives to buy assets in the primary market, which distorts the asset allocation across investors. This distortion of asset allocation further affects the mispricing of assets in the secondary market and in turn affects issuance. In the end, we show that the allocation of assets across investors is inefficient in every decentralized equilibrium.

The inefficiency we find can be interpreted as the result of a double-sided hold-up problem. A hold-up problem, as first described by Williamson (1975) and Klein et al. (1978), arises when one party must bear the entire cost of an investment while others share in the payoff. In markets with trading frictions, hold-up problems arise often because investments must be made ex-ante, before agents meet. For instance, in monetary search models (e.g., Lagos and Wright, 2005; Rocheteau and Wright, 2005; Aruoba et al., 2007) agents acquire money balances before trading with sellers, and in labor search models (e.g. Acemoglu, 1996; Masters, 1998; Acemoglu and Shimer, 1999) firms or workers invest in capital before negotiating wages. In our environment, the hold-up problem is two-sided, faced by both buyers and sellers in the secondary market that must take specific investments before trade. Sellers must invest in newly issued assets in the primary market in order to resell to higher-valuation investors in the secondary market. However, bargaining in the secondary market happens ex-post, after the cost of acquiring the asset is paid. As a result, sellers only internalize a fraction of the social value of reselling the asset because they only receive a share of the gains from trade. To fix this externality, the mechanism would need to assign all the gains from trade to the seller. Likewise, potential buyers in the secondary market must "invest" in not buying assets before trade occurs. When valuing an asset, a buyer understands they have an outside option of waiting to purchase the asset in the future (in either the primary or secondary market). However, ex-post bargaining implies they only internalize a fraction of the social option value of waiting. To fix this externality, the mechanism would need to assign all the gains from trade to the buyer. Both conditions—giving all the bargaining power simultaneously to the seller and the buyer—cannot be jointly satisfied, implying a decentralized-equilibrium allocation that is not constrained-efficient.

Trade is inefficient for any bargaining rule, however the direction of the inefficiency crucially depends on the way buyers and sellers split the surplus in the secondary market. We show that when the secondary-market sellers have all the bargaining power, investors overvalue assets, and issuance and intermediation are inefficiently high; and, when the secondary-market buyers have all the bargaining power, there are little gains to buy newly created assets in order to resell, and issuance and intermediation are inefficiently low.

Our result provides a rational for the types of intervention in OTC markets that were ob-

served during the 2008 financial crisis. This rational is independent of additional frictions, such as private information (see Chang (2014), Chiu and Koeppl (2015), and Bethune et al. (2016), among others). To highlight the role of intervention, we propose a simple government policy that individually corrects the double-sided hold-up problem and decentralizes the constrained efficient solution. Since low-value investors do not fully internalize the gain in intermediating assets, the government subsides their asset holdings. Likewise, since high-value investors do not fully internalize their outside option value of waiting to buy assets in the future, the government taxes their asset holdings. The budget is balanced through lump-sum taxation.

We quantitatively explore the effects of the inefficiency by calibrating our model to match certain features of the US municipal bond market. The municipal market is a textbook example of an active OTC market (see Green et al. (2007), and Bethune et al. (2016) for evidence). The calibration implies that search frictions are minimal—an investor contacts five other investors per day—and that buyers possess a high bargaining power. As a result, the incentives to intermediate assets are too low and aggregate asset supply is suppressed in steady state by 10%. The intervention stimulates intermediation and implies an increase in the bid-ask spread in steady state from 1.67% (a target in the calibration) to 2%. Comparing steady states, welfare increases about 4%; considering the transition from the decentralized steady state, welfare increases about 0.8%.

Finally, we study the effect of an aggregate transitory shock that depresses asset valuations by investors and thus contracts the asset supply in the economy. We designed the shock to mimic the way investors withdrew from asset markets during the financial crisis. We find that the response of the economy with no intervention is more severe than the response of the economy with intervention in two aspects. Welfare losses in the economy with no intervention are larger than in the economy with intervention, and the recovery of the economy with no intervention lags the recovery of the economy with intervention by about one year. This suggests that policies that correct trade inefficiencies natural to OTC markets can make recessions and crisis shorter and milder.

The paper is structured as follows. Section 2 introduces the environment. Section 3 defines a decentralized equilibrium and discusses how bargaining determines equilibrium asset allocations. Section 4 describes the constrained efficient benchmark and compares it with the decentralized equilibrium. Section 5 discuss a government intervention to decentralize the constrained efficient outcome. Section 6 provides a numerical exploration of the model. Finally, Section 7 concludes. We provide proofs of all results in the paper in Appendix A.

Literature Following Duffie et al. (2005), the OTC literature has mostly focused on studying trading dynamics in decentralized markets with no meaningful issuance margin and an exogenous supply of assets. For example, Lagos and Rocheteau (2007) and Garleanu (2009) feature unrestricted asset holdings but leave the aggregate asset supply constant. He and Milbradt (2014) include debt maturity but assume that firms reissue assets to replace maturing debt. Recent work has started to include a meaningful role for asset issuance in an OTC setting. Arseneau et al. (2015) examine similar questions as we do in a three-period model in which assets are created in a primary market subject to a costly state verification problem. Alternatively, we characterize equilibrium dynamics in infinite time in which assets are allocated in a frictional primary market. Geromichalos and Herrenbrueck (2016) also consider an environment with asset issuance and decentralized secondary markets, but their focus is on the determination of liquidity and not on efficiency or policy.

Our efficiency results have a similar flavor to those in the OTC literature that introduce some endogenous extensive margin in trade. For instance, Lagos and Rocheteau (2006) consider an environment in which traders' asset holdings are unrestricted and assets are reallocated between traders through a competitive inter-dealer market.[2] They find that under free entry by dealers, the decentralized equilibrium cannot implement the constrained-efficient solution. Efficient entry requires that dealers' bargaining power equal their impact on matching, a la Hosios (1990). However, positive dealer bargaining power leads to a hold-up problem by traders as a result of ex-post bargaining.[3]

Gofman (2014) also considers the efficiency of OTC markets but in an environment in which agents can trade bilaterally according to an exogenous network structure. If the network is complete, in that all traders can trade directly with each other, then the equilibrium is efficient for any set of bargaining powers. However, if the network is incomplete and bargaining powers are strictly between zero and one, a hold-up problem arises. Traders do not internalize the full gain of transferring the asset on valuations further along the network and, as a result, assets may not end up with the highest valuation traders, lowering welfare. Efficiency can be restored for any connected network when sellers possess all the bargaining power. That happens because only sellers face a hold-up problem. We show, however, that introducing asset issuance necessarily introduces an inefficiency that cannot be restored for any set of bargaining powers because both buyers and sellers face a hold-up problem.

Double-sided hold-up problems have been studied in the context of the labor market in which firms and workers make investment decisions before matching and determining

[2] Also see Lagos and Rocheteau (2007) and Lagos and Rocheteau (2009b) for a similar environment.

[3] The hold-up problem also arises in Lagos et al. (2011), who studies dynamic equilibrium in which dealers can hold inventories.

wages. Acemoglu (1996) shows that random search and ex-post bargaining lead to social increasing returns in the production technology that create an externality since the gains from trade must be split. This leads to a similar result that efficiency cannot be restored by choosing the right bargaining power. Masters (1998) shows this type of inefficiency always leads to underinvestment in physical and human capital. Alternatively, in the context of our asset market there could be under or overinvestment since there are not generally increasing returns to investment on both sides of the market. One-side of the market tends to underinvest in assets (sellers in the secondary market) and the other side tends to overinvest (buyers in the secondary market). We show that bargaining power has an important role in determining the shape of inefficiency.

Recent work has highlighted how the presence of intermediaries in OTC markets with random search can sometimes lead to inefficiency. Farboodi et al. (2017) endogenize contact rates in Duffie et al. (2005) and also find that the equilibrium is, in general, inefficient. Traders inefficiently invest in contact rates as a result of a search externality: they do not internalize that increasing their contact rate affects the distribution of other traders' contacts. A Pigouvian tax that charges traders when they make contact and uses the revenue to double the size of the surplus decentralizes the Pareto optimum. Our efficiency result is similar in that the planner would like to increase the size of the surplus in any trade, however the tax/transfer scheme in Farboodi et al. (2017) only achieves the optimum in the case with symmetric bargaining weights. Further, Farboodi et al. (2017) do not consider issuance or endogenous asset supply.

In Farboodi et al. (2016), traders meet randomly but differ with respect to their ability to commit to take-it-or-leave-it offers, and as a result, their bargaining power. If types, or bargaining powers, are exogenous, then equilibrium is efficient since (i) bilateral trade is efficient and (ii) all traders will meet each other, almost surely.[4] However, if bargaining types are endogenous and commitment requires a sunk cost, equilibrium is inefficient.

Also related to our work, Nosal et al. (2014) and Nosal et al. (2016) study an environment with endogenous intermediaries where efficiency only arises if bargaining weights satisfy a version Hosios (1990) condition.

[4]Related to Gofman (2014), in a random search environment with a fixed supply of assets, the trading network is almost surely complete and equilibrium is efficient even if bargaining powers are inside the unit interval.

2 Environment

Time is continuous and goes from zero to infinity. There are two types of agents: issuers and investors. All agents are infinitely lived, risk-neutral, and discount the future at rate $r > 0$. Investors have measure two and issuers have measure one. These choice of measures simplify the notation but are not necessary for our results. There are objects called assets that issuers issue (or produce) and investors value for their flow of a good called dividends. An asset matures with Poisson arrival rate $\mu > 0$ and pays a unit flow of dividends until maturity. The asset disappears upon maturity and has no terminal payment. Dividends are not tradable and investors must hold an asset to consume its dividends.

Each issuer has to pay an issuance cost to issue an asset, and they do not value asset dividends. The issuance cost is heterogeneous among issuers. It follows an uniform distribution, with density $g(c) = \frac{1}{\bar{c}-\underline{c}}$ and cumulative distribution $G(c) = \frac{c-\underline{c}}{\bar{c}-\underline{c}}$ in the interval $[\underline{c}, \bar{c}]$. Since issuers do not value dividends, they do not hold assets in equilibrium.

Investors are one of two types, low or high, and types are fixed over time.[5] Half of the investors are low type and the remaining half are high type. An investor's type is associated with their utility from consuming dividends. Low-type investors have utility $\nu_l \geq 0$, while high-type investors have utility $\nu_h > \nu_l$. Investors' asset holdings are discrete, either zero or one. We call investors holding an asset *owners* and those not holding an asset *non owners*. Let $\phi_l^0 \in [0,1]$ and $\phi_h^0 \in [0,1]$ denote the measures of low and high-type owner investors that are given in period $t = 0$.

Assumption 1. *(i)* $\nu_l/(r+\mu) = \underline{c}$ *, and (ii)* $\nu_h/(r+\mu) \leq \bar{c}$.

We impose Assumption 1 throughout the paper. Part (i) implies that low-type investors only hold assets if they profit from reselling the asset in the secondary market. To see this, notice that $\nu_l/(r+\mu)$ is the discounted present value that a low-type investor would obtain from holding an asset until maturity. Because this discounted flow equals the lowest production cost, $\underline{c}$, low-type investors never find it profitable to buy the asset only to consume its dividends until maturity. This assumption permits us to isolate the intermediation channel in the model and thus allows us to study how the gains from intermediation shape asset issuance and allocations. Part (ii) guarantees that the allocation is interior, which allows us to take derivatives when needed.

[5] Much of the OTC literature following Duffie et al. (2005) and Lagos and Rocheteau (2007) uses preference shocks to generate trade motives in the secondary market. In our setting, asset maturity already generates trade motives and preference shocks do not add to our analysis. For this reason, we eliminate preference shocks from most of the paper. The exceptions are the numerical exercises of section 6, where the model is augmented to include preference shocks in order to improve the model's fit of the data.

Issuers contact investors at random with Poisson arrival rate $2\lambda_p > 0$, and investors contact other investors at random with Poisson arrival rate $\lambda_s/2 > 0$. Note that an investor expects to meet with an issuer at Poisson arrival rate λ_p; an issuer contacts an investor at rate $2\lambda_p$, and those contacts are distributed among the measure two of investors. Note also that an investor expects to meet with another investor at Poisson arrival rate λ_s; at rate $\lambda_s/2$ he contacts another investor, and at rate $\lambda_s/2$ another investor contacts him. We call the market in which issuers sell newly issued assets to investors *the primary market*, and the market in which investors sell existing assets to each other *the secondary market*. Meetings are bilateral, and utility is transferable across investors and issuers.

To facilitate the analysis here, it is useful to anticipate some equilibrium trading patterns. First, low-type investors sell assets to high-type investors and never the other way around. Thus, low-type investors act as intermediaries. Second, if an issuer with cost c issues to a particular investor—either of high or low type—then all issuers with cost below c also issue to this particular investor. As a result, when defining an allocation, we define the issuance policy only in terms of issuance thresholds c_l and c_h. We later verify that these restrictions are without loss of generality in terms of characterizing decentralized equilibria and efficient allocations in our economy.

An asset allocation and issuance policy define an allocation of the economy. An asset allocation is a map $\boldsymbol{\phi} = \{\phi_l(t), \phi_h(t)\}_t$, where $\phi_l(t)$ and $\phi_h(t)$ denote the measure of low- and high-type owner investors at time t. An issuance policy is a map $\boldsymbol{c} = \{c_l(t), c_h(t)\}_t$, where $c_l(t)$ and $c_h(t)$ denote thresholds for the issuance cost such that issuers with cost below $c_l(t)$ and $c_h(t)$ issue to low- and high-type non owner investors in time t meetings. We drop the argument t from an allocation in order to keep the notation simple whenever it does not cause confusion.

An issuance policy $\boldsymbol{c}$ implements an asset allocation $\boldsymbol{\phi}$ if the measures of owner investors, ϕ_l and ϕ_h, are consistent with the issuance thresholds, c_l and c_h, the trade pattern in the secondary market, and the initial asset holdings ϕ_l^0 and ϕ_h^0. That is, the law of motion for ϕ_l and ϕ_h solve the system of differential equations

$$\dot{\phi}_l = \lambda_p G(c_l)(1-\phi_l) - \mu\phi_l - \lambda_s\phi_l(1-\phi_h) \quad \text{and} \tag{1}$$

$$\dot{\phi}_h = \lambda_p G(c_h)(1-\phi_h) - \mu\phi_h + \lambda_s\phi_l(1-\phi_h), \tag{2}$$

given the initial conditions $\phi_l(0) = \phi_l^0$ and $\phi_h(0) = \phi_h^0$. We say that an asset allocation $\boldsymbol{\phi}$ is feasible if there is an issuance policy $\boldsymbol{c}$ that implements $\boldsymbol{\phi}$.

Equation (1) describes the law of motion for low-type owner investors at time t. On the right hand side, the first term captures the inflow of low-type non owner investors that

meet issuers with cost below c_l and thus buy assets from these issuers. The second term captures the outflow of low-type owners due to asset maturity. The third term captures the outflow of low-type owners due to the fact that they meet high-type non owners and sell their assets to them. The law of motion for high-type owner investors described in (2) follows similarly, except that the last term represents the inflow from trade between investors. This difference between the asset flow equations for low- and high-type investors follows from the different roles played by both types of investors in the secondary market: low-type investors are sellers, while high-type investors are buyers.

3 Decentralized equilibrium

We characterize a decentralized equilibrium in terms of investors' reservation value. The reservation value is the absolute change in value function—that is, in expected present value of utility—induced by acquiring or giving up an asset. Let $V_l(q)$ denote the value function of low-type investors, where $q \in \{0, 1\}$ denotes the asset holdings of the investor. Then $\Delta_l = V_l(1) - V_l(0)$ is the reservation value of low-type investors. Analogously, let $V_h(q)$ denote the value function of high-type investors. Then $\Delta_h = V_h(1) - V_h(0)$ is the reservation value of high-type investors.

Nash bargaining determines trade and price in the primary market, where we assume that buyers—the investors—have all the bargaining power. We make this assumption for two reasons. First, abstracting from general equilibrium effects, the fact that buyers have full bargaining power makes trade in the primary market efficient. Second, and more importantly, the assumption simplifies the environment, and has no qualitative implications for our results regarding the inefficiency of decentralized equilibrium, as analyzed in Section 4. Low- and high-type non owner investors buy an asset if their reservation value is greater than the issuer's issuance cost. Because buyers hold all the bargaining power, the price equals the issuance cost. These trade outcomes imply the trade pattern in the primary market we anticipated earlier. Namely, if an issuer with cost c issues to a particular investor, either high or low type, then all issuers with cost below c must also issue to this particular investor. The reservation values, Δ_l and Δ_h, specify the thresholds for asset issuance.

Nash bargaining also determines trade and price in the secondary market, where buyers have bargaining power $\theta \in [0, 1]$ and sellers have bargaining power $1 - \theta$. To maximize the surplus in a trade, as required by Nash bargaining, the investor with the higher reservation value buys the asset from the investor with the lower reservation value. We anticipate a trade pattern where low-type investors sell assets to high-type investors, as discussed

earlier in the paper. To obtain this pattern, we conjecture that the reservation value of high-type investors is strictly higher than the one of low-type investors—that is $\Delta_h > \Delta_l$. We verify this inequality later using equilibrium equations. Denote by x the price a non owner high-type investor pays to buy an asset from a low-type owner investor. By Nash bargaining, x maximizes $(x - \Delta_l)^{1-\theta}(\Delta_h - x)^{\theta}$, which implies $x = \Delta_h - \theta[\Delta_h - \Delta_l]$.

Given these outcomes in bilateral trade, the value functions for low-type owner and non owner investors satisfy

$$rV_l(0) = \dot{V}_l(0) + \lambda_p \int_{\underline{c}}^{\Delta_l} (\Delta_l - c)g(c)dc \quad \text{and} \tag{3}$$

$$rV_l(1) = \dot{V}_l(1) + \nu_l - \mu\Delta_l + \lambda_s(1 - \phi_h)(1 - \theta)(\Delta_h - \Delta_l). \tag{4}$$

Equation (3) describes the law of motion for the value function of low-type non owner investors. The first term is the change in utility at a point in time and the second term is the expected gain in utility from meeting and purchasing an asset from an issuer in the primary market. Equation (4) describes the law of motion for the value function of low-type owner investors. The first term is the change in utility at a point in time, the second term is the utility flow from holding the asset, the third term is the expected loss in utility due to asset maturity, and the last term is the gain in utility from meeting and selling an asset to a high-type non owner investor in the secondary market.

Similarly, the value function of high-type owner and non owner investors satisfy

$$rV_h(0) = \dot{V}_h(0) + \lambda_p \int_{\underline{c}}^{\Delta_h} (\Delta_h - c)g(c)dc + \lambda_s\phi_l\theta(\Delta_h - \Delta_l) \quad \text{and} \tag{5}$$

$$rV_h(1) = \dot{V}_h(1) + \nu_h - \mu\Delta_h. \tag{6}$$

Equation (5) describes the law of motion for the value function of high-type non owner investors. The first term is the change in utility at a point in time, the second term is the expected gain in utility from meeting and purchasing an asset from an issuer in the primary market, and the last term is the expected gain in utility from meeting and purchasing an asset from a low-type owner investor in the secondary market. Equation (6) describes the law of motion for the value function of high-type owner investors. The first term is the change in utility at a point in time, the second term is the utility flow from holding the asset, and the last term is the expected loss in utility due to asset maturity.

We obtain a differential equation for the reservation value of low-type investors by taking the difference between equations (3) and (4), and for the reservation value of high-type investors by taking the difference between equations (5) and (6). The reservation

values of low- and high-type investors satisfy

$$(r+\mu)\Delta_l = \dot{\Delta}_l + \nu_l - \lambda_p \int_{\underline{c}}^{\Delta_l} (\Delta_l - c)g(c)dc + \lambda_s(1-\phi_h)(1-\theta)(\Delta_h - \Delta_l) \quad \text{and} \tag{7}$$

$$(r+\mu)\Delta_h = \dot{\Delta}_h + \nu_h - \lambda_p \int_{\underline{c}}^{\Delta_h} (\Delta_h - c)g(c)dc - \lambda_s\phi_l\theta(\Delta_h - \Delta_l). \tag{8}$$

Investors' reservation values can be decomposed into three components: a fundamental value and two option values. To understand this decomposition, consider the reservation value equations (7) and (8) in steady state:

$$\Delta_l = \frac{\nu_l}{r+\mu} - \lambda_p \frac{\int_{\underline{c}}^{\Delta_l}(\Delta_l - c)g(c)dc}{r+\mu} + \lambda_s \frac{(1-\phi_h)(1-\theta)(\Delta_h - \Delta_l)}{r+\mu} \quad \text{and}$$

$$\Delta_h = \frac{\nu_h}{r+\mu} - \lambda_p \frac{\int_{\underline{c}}^{\Delta_h}(\Delta_h - c)g(c)dc}{r+\mu} - \lambda_s \frac{\phi_l\theta(\Delta_h - \Delta_l)}{r+\mu}.$$

The first term in the first equation represents the fundamental value of the asset to low-type investors—the expected discounted utility flow from consuming the dividend, ν_l, until maturity. If there was no trade (for instance, if λ_p and λ_s were equal to zero), the fundamental value alone would determine the reservation value Δ_l. However, since there is trade, the options of buying and selling assets in the future play an important role in determining the reservation value. Consider the reservation value for low-type investors. When purchasing an asset, they lose the option of waiting to purchase the asset later in the primary market but gain the option of selling the asset in the secondary market to a high-type investor. These are represented as the second and third terms in the first equation above. Similarly, when high-type investors purchase an asset, they lose the option of waiting to purchase the asset later in the primary market and also lose the option of waiting to purchase the asset later in the secondary market. These are represented as the second and third terms in the second equation above.

Using (7) and (8), Lemma 1 verifies that the reservation value of high-type investors is strictly higher than the reservation value of low-type investors—that is $\Delta_h > \Delta_l$.

Lemma 1. *If reservation values $\Delta_l(t)$ and $\Delta_h(t)$ are bounded and satisfy the differential equations (7) and (8), the reservation value of high-type investors is strictly higher than the reservation value of low-type investors at any point in time. That is, $\Delta_h(t) > \Delta_l(t)$ for all t.*

High-type investors enjoy a higher flow utility from consuming dividends, $\nu_h > \nu_l$. As a result, the fundamental-value component of the reservation value for the high-type investors is higher than that for the low-type investors. If reservation values are not consistent with

the fundamental component, that is if $\Delta_h \leq \Delta_l$, then they must be consistent with beliefs about future reservation values. Low-type investors must believe that their reservation value will keep increasing. This generates an explosive path for the reservation value that is inconsistent with bounded payoffs, a contradiction that implies that reservation values are consistent with their fundamental-value component.[6]

Without loss of generality, we define a decentralized equilibrium in terms of reservation values instead of value functions. For any pair of reservation values, the differential equations (3)-(6) yield the associated value functions. Additionally, we do not include the value function of issuers in our equilibrium definition since it is always zero because they have no bargaining power and are paid their cost when issuing an asset.

Definition 1. *A decentralized equilibrium is an asset allocation and bounded reservation values for investors,* $\{\boldsymbol{\phi}, \boldsymbol{\Delta}\} = \{\phi_l, \phi_h, \Delta_l, \Delta_h\}$*, that solve the differential equations*

$$(r+\mu)\Delta_l = \dot{\Delta}_l + \nu_l - \lambda_p \int_{\underline{c}}^{\Delta_l} (\Delta_l - c) g(c) dc + \lambda_s (1-\phi_h)(1-\theta)(\Delta_h - \Delta_l) \tag{9}$$

$$\dot{\phi}_l = \lambda_p (1-\phi_l) G(\Delta_l) - \mu \phi_l - \lambda_s \phi_l (1-\phi_h) \tag{10}$$

$$(r+\mu)\Delta_h = \dot{\Delta}_h + \nu_h - \lambda_p \int_{\underline{c}}^{\Delta_h} (\Delta_h - c) g(c) dc - \lambda_s \phi_l \theta (\Delta_h - \Delta_l) \tag{11}$$

$$\dot{\phi}_h = \lambda_p (1-\phi_h) G(\Delta_h) - \mu \phi_h + \lambda_s \phi_l (1-\phi_h) \tag{12}$$

with initial conditions $\phi_l(0) = \phi_l^0$ *and* $\phi_h(0) = \phi_h^0$.

A decentralized equilibrium is at a steady state when the asset allocation and reservation value of investors are constant.

Definition 2. *A decentralized steady-state equilibrium is an asset allocation and bounded reservation values for dealers and investors,* $\{\boldsymbol{\phi}, \boldsymbol{\Delta}\} = \{\phi_l, \phi_h, \Delta_l, \Delta_h\}$*, that solve the system of equations* (9)-(12) *with time derivative* $\dot{\phi}_l = \dot{\phi}_h = \dot{\Delta}_l = \dot{\Delta}_h = 0$ *for all* t.

The existence of a decentralized equilibrium (in and out of steady state) follows from standard methods used to solve non linear differential equations.[7]

[6] Lemma 1 confirms our assumption that $\Delta_h > \Delta_l$; however, it does not show that an alternative equilibrium with $\Delta_h \leq \Delta_l$ does not exist. That is, if we assume $\Delta_h \leq \Delta_l$, the reservation value equations (7) and (8) would be different because Nash bargaining would imply trade in the opposite direction—-low-type investors buying assets from high-type investors. These new equations could be consistent with $\Delta_h \leq \Delta_l$, however we show that this is not the case. The trade pattern associated with $\Delta_h < \Delta_l$ implies that $\Delta_h > \Delta_l$, which is a contradiction, and $\Delta_h = \Delta_l$ can easily be ruled out because $\nu_l \neq \nu_h$. We discuss these results in the Appendix with the proof of Lemma 1.

[7] We omit an existence proof here but provide it upon request.

3.1 Bargaining power and equilibrium asset allocations

Bargaining power in the secondary market plays an important role in shaping the incentives of investors to buy or sell assets, which makes it key in determining equilibrium asset allocations. This is not clear from the existing literature. For example, in the standard DGP model, the equilibrium asset allocation is independent of the bargaining power—which determines only transfers of utils. In our model, the equilibrium asset allocation is a function of the bargaining power. In this section, we study this function.

We compare three state-steady economies: the sellers' economy, the buyers' economy, and the interior economy. The three economies are the same in all primitives except for the bargaining power of buyers and sellers in the secondary market. In the sellers' economy, sellers have all the bargaining power, that is $\theta = 0$. In the buyers' economy, buyers have all the bargaining power, that is $\theta = 1$. In the interior economy, neither sellers or buyers have all the bargaining power, that is $0 < \theta < 1$. We order the economies from A to C according to the bargaining power, that is $0 = \theta^A < \theta^B < \theta^C = 1$.

Proposition 1. *If the asset allocations and reservation values* $\{\phi_l^A, \phi_h^A, \Delta_l^A, \Delta_h^A\}$, $\{\phi_l^B, \phi_h^B, \Delta_l^B, \Delta_h^B\}$, *and* $\{\phi_l^C, \phi_h^C, \Delta_l^C, \Delta_h^C\}$ *are associated with steady-state decentralized equilibria for the sellers' economy, the interior economy, and the buyers' economy, then*

(i) $\phi_l^A > \phi_l^B > \phi_l^C = 0$,

(ii) $\phi_h^A > \phi_h^B > \phi_h^C$,

(iii) $\Delta_l^A > \Delta_l^B > \Delta_l^C = \frac{\nu_l}{\mu + r}$, *and*

(iv) $\Delta_h^A = \Delta_h^C > \Delta_h^B$.

Part (i) and (ii) of Proposition 1 provide that low- and high-type investors hold the least amount of assets when buyers possess all the bargaining power and hold the most when sellers possess all the bargaining power. An immediate implication is that aggregate asset supply, $\phi_l + \phi_h$, follows a similar pattern. These results follow by noticing that asset holdings move in the same direction as the reservation value of low-type investors (part (iii) of the proposition). Low-type investors serve as natural intermediates—buying assets from issuers and selling them to high-type investors. In the buyers' economy, low-type investors have a low reservation value because they do not get any of the gains from trade when selling assets. As a result, low-type investors have no incentive to intermediate asset issuance and hold no assets, which reduces the buying options of high-type investors and leads them to hold less assets in equilibrium. The sellers' economy features the opposite pattern. Low-type investors have a high reservation value because they correctly size all the

gains from trade when selling assets. As a result, low-type investors have higher incentives to intermediate assets, increasing the option value of buying assets for high-type investors and leading them to hold more assets in equilibrium. Finally, part (iv) of the proposition provides that the reservation value of high-type investors is maximized for bargaining powers in the extremes, $\theta \in \{0, 1\}$. This follows because when $\theta = 0$ or $\theta = 1$ the option value of buying seasoned assets in the secondary market is zero for high-type investors, and positive option values reduce the value of holding, and thus acquiring, an asset. When $\theta = 0$, the option value is zero because all of the gains from trade are captured by the seller—the low-type investor. When $\theta = 1$, the option value is zero because there is no trade of seasoned assets as low-type investors have no incentives to intermediate.

4 Efficient asset allocation

We now turn to solving for the constrained-efficient allocation, constrained in the sense that the allocation has to satisfy the search frictions of the economy. We label the problem of finding a constrained-efficient asset allocation the planner's problem. The planner is allowed to choose the issuance policy in the primary market and to choose the trade pattern in the secondary market. As we did with the decentralized equilibrium, we anticipate a few results. First, when an owner low-type investor meets a non owner high-type investor, the planner will transfer the asset from the low-type investor to the high-type investor. Second, when an owner high-type investor meets a non owner low-type investor, the planner will not transfer the asset from the high-type investor to the low-type investor.

An asset allocation, $\boldsymbol{\phi}$, is constrained-efficient if there is an issuance policy, $\boldsymbol{c}$, such that $\boldsymbol{\phi}$ and $\boldsymbol{c}$ maximize aggregate utility,

$$\int_0^\infty e^{-rt} \left\{ \phi_l \nu_l + \phi_h \nu_h - \lambda_p \left[(1 - \phi_l) \int_0^{c_l} c g(c) dc + (1 - \phi_h) \int_0^{c_h} c g(c) dc \right] \right\} dt, \tag{13}$$

subject to the feasibility conditions (1) and (2).[8] Lemma 2 provides the first order conditions that are necessary for an asset allocation to be constrained-efficient.

Lemma 2. *If an asset allocation $\boldsymbol{\phi}$ is constrained-efficient, then there exist co-state variables γ_l, $\gamma_h \geq 0$ such that γ_l, γ_h, ϕ_l, and ϕ_h solve the system of differential equations given by*

$$r\gamma_l = \dot{\gamma}_l + \nu_l - \mu\gamma_l - \lambda_p \int_0^{\gamma_l} (\gamma_l - c) g(c) dc + \lambda_s (1 - \phi_h)(\gamma_h - \gamma_l) \tag{14}$$

[8] Note that in this definition we give Pareto weight one to each agent. This is without loss of generality because the model has transferable utility—differences in Pareto weights determine transfers but do not distort the asset allocation.

$$\dot{\phi}_l = \lambda_p(1-\phi_l)G(\gamma_l) - \mu\phi_l - \lambda_s\phi_l(1-\phi_h) \tag{15}$$

$$r\gamma_h = \dot{\gamma}_h + \nu_h - \mu\gamma_h - \lambda_p\int_0^{\gamma_h}(\gamma_h - c)g(c)dc - \lambda_s\phi_l(\gamma_h - \gamma_l) \tag{16}$$

$$\dot{\phi}_h = \lambda_p(1-\phi_h)G(\gamma_h) - \mu\phi_h + \lambda_s\phi_l(1-\phi_h) \tag{17}$$

with boundary conditions $\phi_l(0) = \phi_l^0$, $\phi_h(0) = \phi_h^0$, *and* $\lim e^{-rt}\gamma_l = \lim e^{-rt}\gamma_h = 0$.

The co-state variables, γ_l and γ_h, represent the social value of having low- and high-type investors holding an asset—we call them the social value of a low- and high-type investor. The social value of an investor is analogous to their reservation value, only from the standpoint of aggregate welfare. First, the planner understands that there is a fundamental value in giving an asset to an investor coming from the dividend valuations, ν_l and ν_h. The planner also understands that there are two option values for each type of investor. For low-type investors, there is the option value of buying the asset later in the primary market, $\lambda_p\int_0^{\gamma_l}(\gamma_l - c)g(c)dc$, which he loses by acquiring an asset, and the option value of transferring the asset later to a high-type investor, $\lambda_s(1-\phi_h)(\gamma_h - \gamma_l)$, which he gains by acquiring an asset. For high-type investors, there is the option value of buying the asset later in the primary market, $\lambda_p\int_0^{\gamma_h}(\gamma_h - c)g(c)dc$, which he loses by acquiring an asset, and the option value of buying the asset later from a low-type investor, $\lambda_s\phi_l(\gamma_h - \gamma_l)$, which he also loses by acquiring an asset.

There is a key difference between the social values in (14) and (16) and private reservation values in (9) and (11). When the planner evaluates the social value of allocating an asset to a low-type investor, he takes into account that the low-type investor will, with some probability, transfer the asset to a high-type investor. In doing so, the planner takes into account the entire surplus generated by transfering an asset from a low-type investor to a high-type investor, or the entire surplus from trade in the secondary market. When the low-type investor evaluates his reservation value, however, he takes into account only a fraction $1-\theta$, his bargaining power, of this surplus. The reason is because he faces a hold-up problem—his decision to invest in buying an asset occurs before meeting with a buyer for the asset in the secondary market. As usual in hold-up problems, the only way that the equations for the social value and reservation value of a low-type investor coincide is if the low-type investor has all the bargaining power when selling an asset. That is, $1-\theta = 1$.

Similarly, when the planner evaluates the social value of a high-type investor, he takes into account that the high-type investor will, with some probability, meet with a low-type owner investor in the future in which he could have bought the asset from. In doing so, the planner takes into account the entire loss of surplus generated by passing an asset from a low-type investor to a high-type investor. When the high-type investor evaluates

his reservation value, he takes into account only a fraction θ, his bargaining power, of this surplus. The high-type investors also faces a hold-up problem—his decision not to invest in buying an asset occurs before meeting with a seller in the secondary market. The only way that the equations for the social value and reservation value of a high-type investor coincide is if the high-type investor has all the bargaining power when buying an asset. That is, $\theta = 1$.

Making sense of the two sides of the hold-up problem is not possible, as this would require that both buyers and sellers have all the surplus generated by a trade. Investors will never value the gains from trade in the same way the planner does, and as a result, the outcome of a decentralized equilibrium cannot replicate the planner's solution—no matter how investors bargain over gains from trade. We prove Proposition 2 with a formal version of this argument.

Proposition 2. *A decentralized-equilibrium asset allocation is never efficient.*

The way in which investors split the surplus when trading is irrelevant in concluding that decentralized trade is inefficient. For any surplus splitting rule, investors cannot both fully internalize the social value of trade that leads to inefficiency. However, the surplus splitting rule does matter in determining the direction the equilibrium allocation is distorted. To illustrate this, consider again the sellers' and buyers' economy associated with bargaining powers $\theta^A = 0$ and $\theta^C = 1$ from Section 3.1. In the following proposition, we show how the bargaining power determines equilibrium allocations.

Proposition 3. *If the asset allocation and reservation values $\{\phi_l^A, \phi_h^A, \Delta_l^A, \Delta_h^A\}$ are associated with a steady-state decentralized equilibria for the sellers' economy, $\{\phi_l^C, \phi_h^C, \Delta_l^C, \Delta_h^C\}$ are associated with a steady-state decentralized equilibria for the buyers' economy, and the asset allocation (ϕ_l^*, ϕ_h^*) is efficient, in steady state, and has social values γ_l and γ_h, then*

(i) $\phi_l^A > \phi_l^* > \phi_l^C = 0$,

(ii) $\phi_h^A > \phi_h^* > \phi_h^C$,

(iii) $\Delta_l^A > \gamma_l > \Delta_l^C = \frac{\nu_l}{\mu+r}$, *and*

(iv) $\Delta_h^A = \Delta_h^C > \gamma_h$.

In the sellers' economy, high-type investors do not internalize the option value of buying assets in the secondary market because they gain no surplus. As a result, they over value purchasing assets through the primary market and $\Delta_h^A > \gamma_h$. Low-type investors also over value assets since the secondary market surplus reflects the over valuation of high-type

investors, $\Delta_l^A > \gamma_l$. In equilibrium, there is over issuance and the asset supply is too high, $\phi_l^A + \phi_h^A > \phi_l^* + \phi_h^*$. The opposite is true in the buyers' economy. In this case, low-type investors do not gain any surplus from reselling and, since $\underline{c} = \nu_l/(r+\mu)$, they do not hold assets, $\phi_l^C = 0$ and $\Delta_l^C < \gamma_l$. Since there is no secondary market trade, high-type investors have no secondary market option value, which again implies they over value primary issuance, $\Delta_h^C > \gamma_h$. In equilibrium, there is under issuance and the asset supply is too low, $\phi_l^C + \phi_h^C < \phi_l^* + \phi_h^*$.

Two comments are relevant before concluding this section. The first comment is that the inefficiency result is independent of the type of bilateral bargaining protocol that we use in the model. What is key for the result is that buyers and sellers in a meeting need to make costly actions prior to the meeting, thus allowing a double hold-up problem to arise.[9]

The second comment pertains the way we split the surplus in the primary market. We assigned full bargaining power to buyers in the primary market for two reasons. Assigning the full bargaining power to the buyer is optimal if we abstract from the secondary market, and the inefficiency result survives even if we allow the splitting rule in the primary market to differ. Because of these reasons, we decided to abstract from having a generic splitting rule in the primary market to keep the model as simple as possible. This allows us to focus on how the way trade surpluses are split in the secondary market between buyers and sellers affects the economy and generates the inefficiency we discuss in Proposition 2. Still, a proof of the inefficiency result with a generic splitting rule in the primary market is available for the interested reader.

5 A government intervention

Decentralized trade in the secondary market with ex-post bargaining necessarily leads to wedges between the investors' reservation values and the social values coming from investors not internalizing the full surplus generated by their trade. To correct for these wedges, we propose a simple tax-subsidy government intervention. Since low-type investors, when they buy an asset, do not internalize the full gain in the option value of selling later in the secondary market, we propose subsidizing their asset holdings. Since

[9] Note, however, that a more general mechanism or trading protocol could, in principle, approximate the efficient outcome. For instance, if trades and transfers are history dependent, efficiency could arise due to a form of folk theorem which holds in this environment. Alternatively, price posting and directed search have been shown to solve double-sided hold-up problems in environments with constant returns to scale matching (which is not true in the benchmark OTC environment we use). We do not allow for more general mechanisms or trading protocols because we interpret these markets as spot trading and take seriously the notion that investors are limited in their information about potential trading partners, as is broadly considered the case in the OTC literature.

high-type investors, when they buy an asset, do not internalize the full loss in the option value of buying later in the secondary market, we propose a tax to their asset holdings. We achieve a balanced budget through lump-sum taxation. We show that this simple policy fully corrects for the double hold-up problem and achieves efficiency.

Formally, a government intervention, or just an intervention to keep it simple, is a triple $\boldsymbol{\tau} = \{\tau_l(t), \tau_h(t), \bar{\tau}(t)\}_t$, where τ_l is a subsidy on asset holdings of low-type investors, τ_h is a tax on asset holdings of high-type investors, and $\bar{\tau}$ is a lump-sum tax on all investors. Given an asset allocation $\boldsymbol{\phi}$, an intervention $\boldsymbol{\tau} = \{\tau_l(t), \tau_h(t), \bar{\tau}(t)\}_t$ is feasible if

$$\int_0^\infty e^{-rt}[2\bar{\tau}(t) + \phi_h\tau_h(t) - \phi_l\tau_l(t)]dt \geq 0. \tag{18}$$

We adjust the decentralized equilibrium equations to account for the intervention and get the following equilibrium definition.

Definition 3. *A decentralized equilibrium with intervention $\boldsymbol{\tau}$ is an asset allocation and bounded reservation values for investors, $\{\boldsymbol{\phi}, \boldsymbol{\Delta}\} = \{\phi_l, \phi_h, \Delta_l, \Delta_h\}$, that solve the equations*

$$(r+\mu)\Delta_l = \dot{\Delta}_l + \nu_l + \tau_l - \lambda_p \int_{\underline{c}}^{\Delta_l} (\Delta_l - c)g(c)dc + \lambda_s(1-\phi_h)(1-\theta)(\Delta_h - \Delta_l) \tag{19}$$

$$\dot{\phi}_l = \lambda_p(1-\phi_l)G(\Delta_l) - \mu\phi_l - \lambda_s\phi_l(1-\phi_h) \tag{20}$$

$$(r+\mu)\Delta_h = \dot{\Delta}_h + \nu_h - \tau_h - \lambda_p \int_{\underline{c}}^{\Delta_h} (\Delta_h - c)g(c)dc - \lambda_s\phi_l\theta(\Delta_h - \Delta_l) \tag{21}$$

$$\dot{\phi}_h = \lambda_p(1-\phi_h)G(\Delta_h) - \mu\phi_h + \lambda_s\phi_l(1-\phi_h) \tag{22}$$

$$\int_0^\infty e^{-rt}[2\bar{\tau}(t) + \phi_h\tau_h(t) - \phi_l\tau_l(t)]dt = 0 \tag{23}$$

with initial conditions $\phi_l(0) = \phi_l^0$ and $\phi_h(0) = \phi_h^0$.

Note that the lump-sum tax $\bar{\tau}$ does not appear in the reservation-value equations (19) and (21). This is because investors pay $\bar{\tau}$ both when they are holding or not holding an asset, so $\bar{\tau}$ does not have a direct impact on the gain of holding an asset.

Proposition 4. *Consider an efficient asset allocation, $\boldsymbol{\phi}^*$, associated with bounded social values γ_l and γ_h. Define $\boldsymbol{\tau} = \{\tau_l, \tau_h, \bar{\tau}\}$ as $\tau_l = \theta\lambda_s(1-\phi_h^*)(\gamma_h - \gamma_l)$, $\tau_h = (1-\theta)\lambda_s\phi_l^*(\gamma_h - \gamma_l)$, and $\bar{\tau} = (\phi_l\tau_l - \phi_h\tau_h)/2$. Then $\{\phi_l^*, \phi_h^*, \gamma_l, \gamma_h\}$ is a decentralized equilibrium with intervention $\boldsymbol{\tau}$.*

An immediate implication of Proposition 4 is that the intervention policy simplifies when the bargaining power is either zero or one. If sellers have all the bargaining power ($\theta = 0$), the intervention restores efficiency simply with a tax to asset holdings of high-type

investors. In this case, there is a unique source of inefficiency to be solved: buyers fail to internalize the option value lost when they acquire an asset. Likewise, if buyers have all the bargaining power ($\theta = 1$), the intervention restores efficiency by subsidizing asset holdings of low-type investors. The unique source of inefficiency to be solved is that sellers fail to internalize the option value gained when they acquire an asset. The next corollary formalizes these claims.

Corollary 1. *Consider the buyer's and seller's economy described before. Then, the following holds:*

(i) in the seller's economy, that is $\theta = 0$, if $\boldsymbol{\phi}^*$ is an efficient asset allocation associated with bounded social values γ_l and γ_h, then $\{\phi_l^*, \phi_h^*, \gamma_l, \gamma_h\}$ is a decentralized equilibrium with intervention $\tau_l = 0$, $\tau_h = \lambda_s \phi_l^* (\gamma_h - \gamma_l)$ and $\bar{\tau} = -\phi_h^* \tau_h / 2$; and

(ii) in the buyer's economy, that is $\theta = 1$, if $\boldsymbol{\phi}^*$ is an efficient asset allocation associated with bounded social values γ_l and γ_h, then $\{\phi_l^*, \phi_h^*, \gamma_l, \gamma_h\}$ is a decentralized equilibrium with intervention $\tau_l = \lambda_s (1 - \phi_h^*)(\gamma_h - \gamma_l)$, $\tau_h = 0$ and $\bar{\tau} = \phi_l^* \tau_l / 2$.

6 Numerical exploration of the model

In this section, we numerically explore three features of the model: the inefficiency implied by the double hold-up problem, how the inefficiency interacts with search frictions in the secondary market, and how the economy responds to aggregate shocks, both in autarky and under government intervention.

6.1 Parameters

In our numerical exploration, we intend to work with parameters that are sensible and generate reasonable market outcomes, but we do not claim that our environment fully captures all the features of OTC markets. With this in mind, we set some parameters based on our experience with the model and existing literature, and the remaining parameters we calibrate to match moments of the US municipal bond market. The Municipal Securities Rulemaking Board (MSRB) makes data available covering primary and secondary trade of municipal bonds. The data we use cover the periods between 2005 and 2014. To give an idea of the size of this market, during our sample period, municipalities issued about 108,000 different bonds at a total par value of $3.6 trillion. The secondary market for municipal bonds was also active, with 40,000 transactions per day and $14 million par amount per transaction.[10]

[10] For more information, check the MSRB website http://www.msrb.org.

One issue we found while performing our exercise is that the model generates fewer trades in the secondary market than what we observe in the municipal bond market. The existing literature with a fixed supply of assets features preference shocks in order to generate trade in the secondary market. In order to match the size of the secondary market in our data, we extend our model in this same way. That is, we assume that, with a Poisson arrival rate α, a low-type investor becomes a high-type investor and with the same Poisson arrival rate α, a high-type investor becomes a low-type investor. The parameter α controls the relative size of the secondary market. Everything else constant, high α leads to more trades in the secondary market because when high-type owner investors become low-type owner investors they benefit from selling the asset to high-type non owner investors. In the opposite way, low α leads to less trades in the secondary market. We stress that our theoretical findings survive the introduction of preference shocks. In particular, the inefficiency result provided in Proposition 2 still applies.

With the extension, we have 10 parameters to set: r, ν_l, ν_h, $\underline{c}$, $\bar{c}$, μ, λ_p, λ_s, θ, and α. We set the discount rate $r = 0.05$, which is associated with a time length of one year. We set investors type $\nu_h = 1$ and $\nu_l = 0.75$. We set the issuance cost parameters $\underline{c} = \nu_l/(r+\mu)$ and $\bar{c} = \nu_h/(r+\mu)$. The remaining parameters we calibrate to match moments of the US municipal bond market. We target five moments, average maturity, average yield, bid-ask spread in the secondary market, average time between buying in the primary market and selling in the secondary market, and relative size of the secondary market. In the Appendix, we describe the computation of these moments in the data. Average maturity in the data is 15 years and average maturity in the model is $1/\mu$, so we set $\mu = 1/15$. For λ_p, λ_s, θ, and α, we use a grid search algorithm to minimize the distance between the moments associated with a decentralized steady-state equilibrium and the moments in the data. Table 1 contains the moments we use, and Table 2 contains the final parameters. The model matches relatively well the moments we target, with the exception of the average yield.[11]

Two calibrated parameters stand out and warrant further discussion: the secondary-market search intensity, $\lambda_s = 1618$, and the bargaining power of buyers in the secondary market, $\theta = 0.98$. The secondary-market search intensity, $\lambda_s = 1618$, implies that an investor contacts five other investors per day. This value for λ_s reflects a secondary market of municipal bonds where dealers take only five days to sell an asset—a target we match in

[11] We attempted a version of the calibration in which we included the dividend valuations, ν_l and/or ν_h. These parameters speak directly to prices and, therefore, to yields. When we did that, we added more moments to keep the model identified, and we did not find any big change in what we learn from the numerical exercises so we decided to be parsimonious and keep the number of moments in the paper to a minimum.

Table 1: Moments

	Data	Model
Average yield (%)	5.23	3.97
Bid-ask spread (%)	1.67	1.68
Avg. time to sell (days)	4.99	5.00
Secondary-market share volume (%)	43.87	38.00

Table 2: Parameters

r	μ	λ_p	λ_s	θ	$\underline{c}$	$\bar{c}$	ν_l	ν_h	α
0.05	$1/15$	0.19	1618	0.98	6.43	8.58	0.75	1	0.02

our calibration. The other parameter, which we find more intriguing to understand, is the bargaining power of buyers in the secondary market, $\theta = 0.98$. The bid-ask spread of low-type investors is about 1.7%—which seems high—and still, sellers have only 2% bargaining power. What is going on is that the secondary market allocates assets from low to high-type investors very fast due to the high λ_s. As a result, there are few low-type investors to buy assets from, and the surplus generated in a secondary-market trade, $\Delta_h - \Delta_l$, is high. The bid-ask spread of low-type investors is $\int \frac{\Delta_l + (1-\theta)(\Delta_h - \Delta_l)}{c} dG(c) - 1$. Given that $\Delta_h - \Delta_l$ is high, to get a bid-ask spread of 1.7% we must then have $1 - \theta$ small—that is, we must then have θ close to one.

6.2 The scale and scope of the inefficiency

To evaluate the inefficiency in our economy, we perform the following exercise. We start with an economy that is in a decentralized steady-state equilibrium. Then the planner, in a move that is unexpected by investors, implements the efficient allocation using the policy described in Section 5. Table 3 contains some outcomes from this exercise.

The economy we calibrated is close to the buyers' economy we discussed in previous sections because the bargaining power of buyers in the secondary market is close to one. With the numerical example, we can better see how that works. As in the buyers' economy of Proposition 3, asset holdings are lower for low- and high-type investors than in the constrained-efficient allocation. In the decentralized economy, low-type investors have little incentive to intermediate. The planner increases the bid-ask spread (i.e. the intermediation markup) from 1.67% to 2% in the constrained-efficient allocation. Since the decentralized

Table 3: Outcomes

Steady-state outcomes	Dec. equilibrium	Const.-Efficient
Bid-ask spread (%)	1.67	2.0
Low-type holdings, ϕ_l	0.001	0.078
High-type holdings, ϕ_h	0.96	0.99
Low-type time to sell (days)	4.99	33.8
High-type time to buy (days)	206.6	28.5
Secondary-market tightness (Sellers/Buyers)	0.03	1.19

Welfare gain (%)	Dec. equilibrium	Const.-Efficient
Steady state	–	3.95
Discounted present value	–	0.82

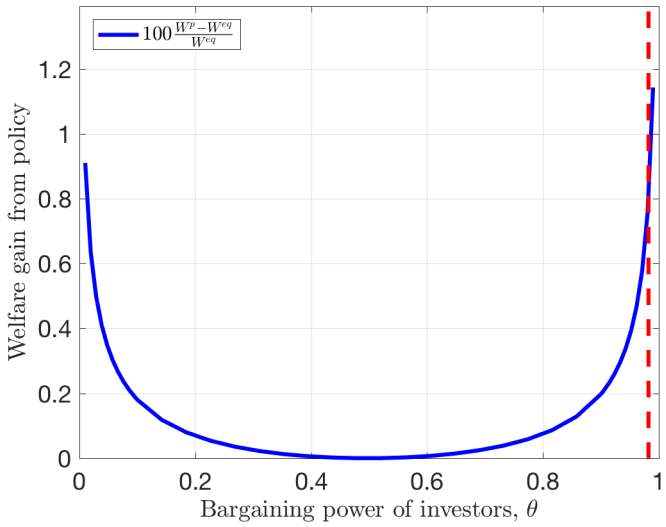
equilibrium features fewer sellers in the secondary market than what is observed in the constrained-efficient allocation, the decentralized equilibrium exhibits lower tightness: 0.03 instead of 1.19 in the efficient allocation. This implies that (i) intermediators sell their assets quickly in the decentralized equilibrium (5 days, on average) and slowly in the constrained-efficient allocation (33 days, on average), and (ii) high-type investors require a long time to buy an asset in the secondary market in the decentralized equilibrium (206 days, on average) while it only takes them a short time to buy assets in the secondary market in the constrained-efficient allocation (28 days, on average). Overall, this implies a 3.95% welfare gain, not including the transition, and 0.82% welfare gain, including the transition of moving from the steady-state decentralized equilibrium to the efficient steady state.

6.3 Effects of bargaining in the secondary market

As described in Proposition 3, the way the surplus is split between investors in the secondary market is key in influencing the direction of misallocation and inefficient asset supply. Figure 1 illustrates the quantitative impact of θ on the magnitude of the efficiency gains from introducing the tax/subsidy scheme. The welfare gains of the policy are highly sensitive to the surplus splitting rule, particularly at the end points. Changing θ from 0.98 in the baseline economy to 1 implies that the welfare gain from the policy increases from 0.82% to about 1.2%. On the other hand, when θ is near one-half, the gains are close to zero.

Why is the magnitude of the inefficiency highly sensitive to θ at the end points? Figure 2 illustrates the steady-state decentralized equilibrium allocation as a function of θ. The

Figure 1: The welfare effects of the optimal tax/subsidy scheme as a function of θ.



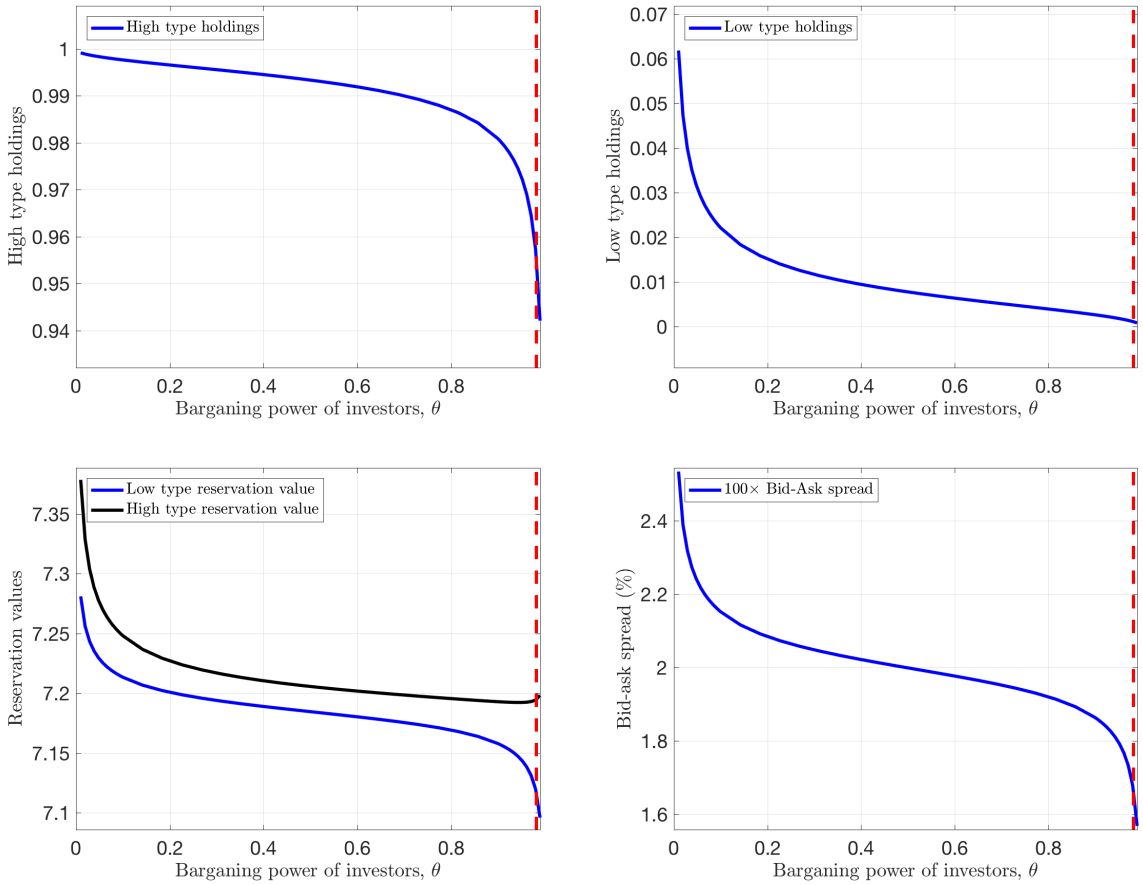
top-two panels show asset holdings of low- and high-type investors, the bottom-left panel shows the reservation values, and the bottom-right panel shows the average bid-ask spread. Small deviations around the calibrated value of θ imply large deviations in the allocation. As θ decreases slightly, the asset holdings of high-type investors increase rapidly. The reservation value of investors converge which implies the gains to secondary market trade fall. For intermediate values of θ, reservation values and the asset allocation are relatively insensitive to changes in θ.

It may seem as though the sensitivity in θ is problematic, for instance perhaps any error in the moments our calibration matches may impact the bargaining power and as a result impact the gain to intervention. However, while the welfare gains are sensitive to θ, its calibrated value is not sensitive to the moments we match. For instance, imposing that $\theta = 1$ would imply a markup of zero and imposing $\theta = 0.97$ would imply a markup of 4.0, both of which are out of line with any of the estimates for dealer markups in Green et al. (2007) (our targets).

6.4 Effects of trading speed

In the baseline economy, secondary market trading speed is high, implying that trading delays are minimal. The municipal bond market does not seem to feature significant frictions in finding a counterparty. Despite this fact, the welfare gain from intervention is still large. Even though search frictions are low, bilateral trade and bargaining still imply that the surplus cannot be fully internalized by both investors. Policies aimed at solely improving trading speed may be limited by the inefficiency of bilateral trade. We highlight these

Figure 2: The decentralized steady-state allocation as a function of θ.

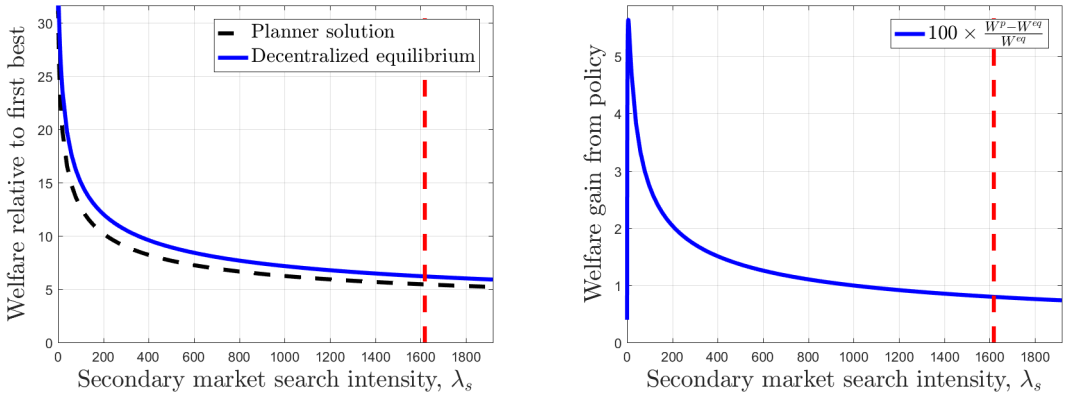


limitations in Figure 3 that illustrates the effect of trading speed on welfare.

The left panel compares welfare in the planner's solution and decentralized equilibrium to the unconstrained, first best allocation. The solid-blue line shows the welfare loss, comparing steady states, of the decentralized equilibrium relative to first-best. Similarly, the dashed-black line shows the welfare loss of the planner's, constrained solution relative to first best. The difference between the two lines represents the gain from intervention, which we plot in the right panel.

As trading speed increases, welfare in both the decentralized equilibrium and the planner's solution increase, causing the welfare loss compared to the first best to decline. For high values of λ_s there is still a considerable welfare difference between the constrained/decentralized allocations and first best. In terms of the gains from intervention, the effects are hump-shaped. When trading speed is zero, there is no scope for intervention—the two lines in the left panel are equal. Markets with a slow trading speed have the greatest need for intervention. However even for high values of λ_s, there is a considerable gain from intervention. Fast markets do not imply that the inefficiency of

Figure 3: The welfare loss relative to unconstrained first best.



decentralized trade is small.

6.5 Application: aggregate demand shocks

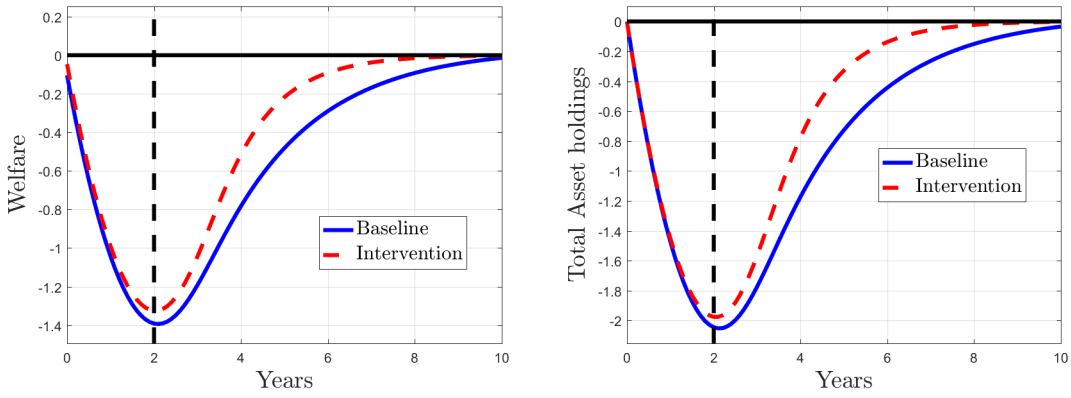
During the 2008/2009 financial crisis, many investors withdrew from the market for asset-backed securities and other OTC asset markets, something we interpret as the result of a negative aggregate demand shock. In this section, we intend to understand how the friction we identify, the double hold-up problem, shapes the response to the shock and how government intervention that restores constrained-efficiency to the economy impacts the severity and duration of the crisis that follows. We view this experiment as providing a good rational for government intervention in OTC markets.

Specifically, we consider an unexpected aggregate shock to the valuation of dividends of investors. Operationally, we assume that the economy is in steady state when an unanticipated shock occurs at time zero that decreases the flow valuation of high-type investors, $\nu_h = 1.00$, to the level of low-type investors, $\nu_l = 0.75$, for $\bar{t}$ years and then slowly recovers. The process $\dot{\nu}_h = (1 - \nu_h)(t - \bar{t})^2$ governs the recovery of ν_h from time $\bar{t}$ forward. Here we set $\bar{t} = 2$. At time zero, after the shock hits the economy, investors and issuers learn the entire future path of ν_h.

We examine how the shock affects asset issuance, welfare, and the distribution of asset holdings in the secondary market in two economies: (i) the decentralized economy we calibrated in Section 6.1 (the "baseline economy"), and (ii) the decentralized economy with taxes that attains the constrained-efficient allocation that we described in Section 5 (the "intervention" economy). Comparing the responses of these two economies to the aggregate demand shock will illustrate that the severity and length of crises in OTC asset markets are lower under government intervention than under no intervention.

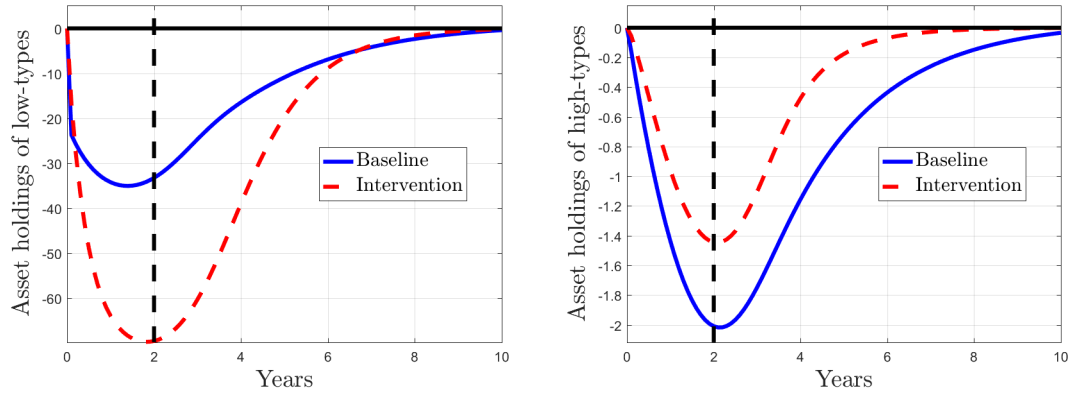
Figure 4 illustrates the responses of welfare and total asset holdings to the aggregate

Figure 4: Response in welfare and asset holdings to an aggregate demand shock.



shock. It shows that the aggregate responses to the shock in the baseline and the intervention economies are similar at a qualitative level but quantitatively different. Under government intervention, welfare and the asset level exhibit a slightly dampened decline during the downturn but a markedly faster recovery than in the baseline economy. At the depth of the crisis, after two years, welfare in the intervention economy is about 1.3% lower than in steady state, while in the baseline economy welfare is 1.4% lower. A similar pattern is present for total asset holdings, a measure that policymakers and pundits often use to gauge welfare, experiencing a 2% fall relative to steady state under government intervention compared to a 2.1% fall in the baseline economy. Both economies start recovering after one years. However, welfare and total asset holdings in the baseline economy lag those in the intervention economy by one year, on average.

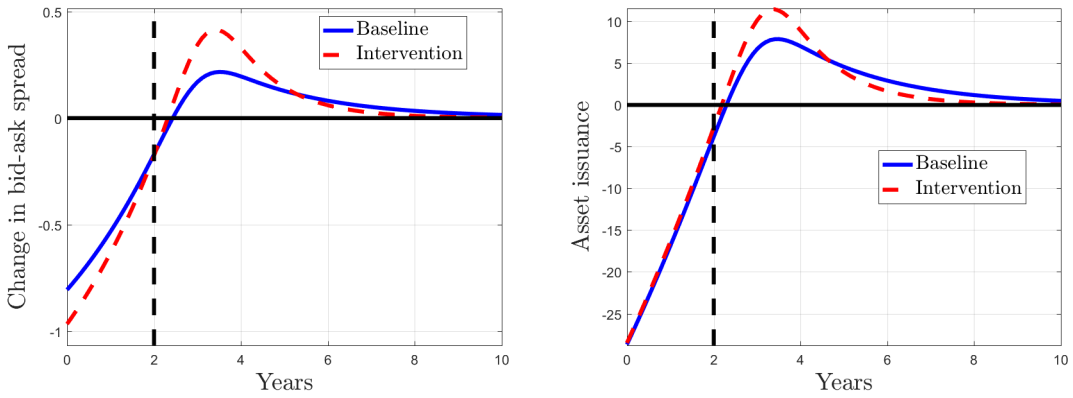
Figure 5: Asset holdings by investor type.



How exactly does the intervention achieve a dampened welfare loss and faster recovery? A notable feature of the intervention economy is that the adjustment largely occurs through a significant fall in the intermediation performed by low-type investors. That is, the intervention induces low-type investors to rapidly respond to demand shocks, either

by depleting their inventories during the downturn or restoring them quickly at the onset of the recovery. This response in intermediation allows the asset holdings of high-type investors to recover faster when their valuation improves. Figure 5 displays the asset holdings of each investor type in order to illustrate this point. The decline in asset holdings of low-type investors (relative to steady-state) is 100% larger in the intervention economy than in the baseline economy. Once the recovery begins (when the high-type investor valuation begins to increase toward its original value) the low-type investors' asset holdings increase quickly, hiking up the intermediation level. In the intervention economy, low-type investors are used as intermediaries to a greater extent in order to reduce the welfare losses that follow from the aggregate demand shock.

Figure 6: Bid-ask spreads and asset issuance.



How does the intervention incentivize low-type investors to act quickly during the demand shock episode? Figure 6 shows the bid-ask spread recieved by low-type investors as well as total asset issuance. Bid-ask spreads generate the correct incentives for low-type investors to adjust their level of intermediation. At the moment the shock hits the economy, bid-ask spreads decline more in the intervention economy than in the baseline economy. This larger decline generates incentives for low-type investors to reduce their intermediation activity more in the intervention economy than in the baseline economy. Once the economy starts to recover, bid-ask spreads increase more rapidly in the intervention economy than in the baseline economy. This generates the correct incentives for low-type investors to increase their intermediation activity more in the intervention economy than in the baseline economy. They do so by soaking up a wave of new asset issuance that is spurred more during the recovery phase in the intervention economy than in the baseline economy. The intervention responds to aggregate demand shocks by using intermediation to dampen and stimulate asset issuance. Doing so leads to a milder downturn and shorter recovery.

7 Conclusion

We showed that the issuance of assets that retrade in OTC secondary markets is always distorted as a result of a double-sided hold-up problem that cannot be solved by correctly splitting the gains from trade. Low-value investors, acting as natural intermediaries, require the full surplus from trade in order to fully internalize the social value of transferring the asset to higher-valuation investors. On the other hand, high-value investors require the full surplus from trade in order to fully internalize their outside option of waiting to purchase the asset in the future. Both conditions can never be simultaneously satisfied, leading to an inefficient aggregate asset supply and misallocation among investors. Further, the direction of inefficiency and misallocation depends on the way the surplus is split. If secondary market buyers posses all the bargaining power, intermediation and asset supply are depressed. If secondary market sellers posses all the bargaining power, then intermediation and asset supply are too high.

The government can restore efficiency by levying a tax on the asset holdings of high-valuation investors and subsidizing asset holdings of low-valuation investors. When calibrating our model to the US municipal bond market, we find that introducing this intervention increases intermediaries' steady-state bid-ask spread from 1.67% to 2%, increases steady state asset supply by 3.1%, and leads to a 4% welfare gain in steady-state. In response to a temporary negative aggregate demand shock, the planner reallocates assets largely through low-valuation (intermediary) investors by decreasing their incentive to intermediate during the downturn and increasing their incentive to intermediate during the recovery. As a result, welfare and the asset level recover considerably faster than in the economy without intervention.

While we chose to illustrate the inefficiency result in a simple environment with two types, discrete asset holdings, and no competitive dealers, we conjecture that these results are quite general. To illustrate this, in Supplementary Appendix C we show the inefficiency holds in an environment with a continuum of investor types as in Hugonnier et al. (2014) or when an investor's bargaining power can vary depending on their trading partner. The key ingredient needed is having a two-sided investment decision (which introducing asset issuance implies) with ex-post bargaining.

References

Acemoglu, D. (1996). A microfoundation for social increasing returns in human capital accumulation. *The Quarterly Journal of Economics*, 111(3):779–804.

Acemoglu, D. and Shimer, R. (1999). Holdups and efficiency with search frictions. *International Economic Review*, 40(4):827–849.

Arseneau, D., Rappoport, D., and Vardoulakis, A. (2015). Secondary market liquidity and the optimal capital structure. Finance and Economics Discussion Series 2015-031, Board of Governors of the Federal Reserve System.

Aruoba, B., Rocheteau, G., and Wright, R. (2007). Bargaining and the value of money. *Journal of Monetary Economics*, 54:2636–2655.

Bethune, Z., Sultanum, B., and Trachter, N. (2016). Private information in over-the-counter markets.

Chang, B. (2014). Adverse selection and liqudity distortion.

Chiu, J. and Koeppl, T. (2015). Trading dynamics with adverse selection and search: Market freeze, intervention and recovery. *Review of Economic Studies*, 00:1–35.

Duffie, D., Garleanu, N., and Pedersen, L. H. (2005). Over-the-counter markets. *Econometrica*, 73:1815–1847.

Duffie, D., Garleanu, N., and Pedersen, L. H. (2007). Valuation in over-the-counter markets. *The Review of Financial Studies*, 20(6):1865–1900.

Farboodi, M., Jarosch, G., and Menzio, G. (2016). Intermediation as rent extraction.

Farboodi, M., Jarosch, G., and Shimer, R. (2017). Meeting technologies in decentralized asset markets.

Garleanu, N. (2009). Portfolio choice and pricing in illiquid markets. *Journal of Economic Theory*, 144:532–564.

Geromichalos, A. and Herrenbrueck, L. (2016). The strategic determination of the supply of liquid assets.

Gofman, M. (2014). A network-based analysis of over-the-counter markets.

Green, R., Hollifield, B., and Schurhoff, N. (2007). Financial intermediation and the costs of trading in an opaque market. *The Review of Financial Studies*, 20(2):275–314.

He, Z. and Milbradt, K. (2014). Endogenous liquidity and defaultable bonds. *Econometrica*, 82(4):1443–1508.

Hosios, A. (1990). On the efficiency of matching and related models of search unemployment. *Review of Economic Studies*, 57:279–298.

Hugonnier, J., Lester, B., and Weill, P.-O. (2014). Hetergeneity in decentralized asset markets.

Klein, B., Crawford, R., and Alchian, A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *The Journal of Law and Economics*, 21(2):297–326.

Lagos, R. and Rocheteau, G. (2006). Search in asset markets. Federal Reserve Bank of Cleveland Working Paper No. 06-07.

Lagos, R. and Rocheteau, G. (2007). Search in asset markets: Market structure, liquidity, and welfare. *American Economic Review: Papers and Proceedings*, 97(2):198–202.

Lagos, R. and Rocheteau, G. (2009a). Liquidity in asset markets with search frictions. *Econometrica*, 77(2):403–426.

Lagos, R. and Rocheteau, G. (2009b). Liquidity in asset markets with search frictions. *Econometrica*, 57:403–426.

Lagos, R., Rocheteau, G., and Weill, P.-O. (2011). Crises and liquidity in over-the-counter markets. *Journal of Economic Theory*, 146(6):2169–2205.

Lagos, R. and Wright, R. (2005). A unified framework for monetary theory and policy analysis. *Journal of Political Economy*, 113(3):463–484.

Masters, A. (1998). Efficiency of investment in human and physical capital in a model of bilateral search and bargaining. *International Economic Review*, 39(2):477–494.

Nosal, E., Wong, Y.-Y., and Wright, R. (2014). More on middlemen: Equilibrium entry and efficiency in intermediated markets. Federal Reserve Bank of Chicago WP 2014-18.

Nosal, E., Wong, Y.-Y., and Wright, R. (2016). Who wants to be a middleman?

Rocheteau, G. and Wright, R. (2005). Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium. *Econometrica*, 73(1):175–202.

Williamson, O. (1975). *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

A Appendix—Proofs

A.1 Proof of Lemma 1

The difference $\Delta_h - \Delta_l$ implies by the reservation-value equations (7) and (8) is

$$\begin{aligned}(r+\mu+2\rho)(\Delta_h-\Delta_l) = (\dot{\Delta}_h-\dot{\Delta}_l) + (\nu_h-\nu_l) \\ - \lambda_p\left[\int_{\underline{c}}^{\Delta_h}(\Delta_h-c)g(c)dc - \int_{\underline{c}}^{\Delta_l}(\Delta_l-c)g(c)dc\right] \\ - \lambda_s\Big[\phi_l\theta + (1-\phi_h)(1-\theta)\Big](\Delta_h-\Delta_l).\end{aligned}$$

We prove the lemma by a contradiction argument. Suppose, by the way of contradiction, that $\Delta_h(t) - \Delta_l(t) \leq 0$ for some time period t. Then the left hand side of the above equation is smaller or equal to zero. However, in the right hand side, all the terms but $\dot{\Delta}_h - \dot{\Delta}_l$ are non-negative, and at least $\nu_h - \nu_l$ is strictly positive. Therefore, $\dot{\Delta}_h - \dot{\Delta}_l \leq -(\nu_h - \nu_l) < 0$ is strictly negative in time t, $\Delta_h - \Delta_l$ stays negative, and $\dot{\Delta}_h - \dot{\Delta}_l \leq -(\nu_h - \nu_l) < 0$ for all $t' \geq t$. But this implies an explosive path for either Δ_h or Δ_l, which contradicts that both functions are bounded. Hence, $\Delta_h(t) - \Delta_l(t) > 0$ for all time periods t.

Further, if we suppose that $\Delta_l > \Delta_h$, and derive the reservation value of low and high type investors associated with the trade pattern implied by $\Delta_l > \Delta_h$, the reservation value equations would be

$$(r+\mu)\Delta_l = \dot{\Delta}_l + \nu_l + \rho(\Delta_h-\Delta_l) - \lambda_p\int_{\underline{c}}^{\Delta_l}(\Delta_l-c)g(c)dc - \lambda_s\phi_h\theta(\Delta_l-\Delta_h) \tag{24}$$

$$(r+\mu)\Delta_h = \dot{\Delta}_h + \nu_h - \rho(\Delta_h-\Delta_l) - \lambda_p\int_{\underline{c}}^{\Delta_h}(\Delta_h-c)g(c)dc + \lambda_s(1-\phi_l)(1-\theta)(\Delta_l-\Delta_h) \tag{25}$$

Which implies that

$$\begin{aligned}(r+\mu+2\rho)(\Delta_h-\Delta_l) = (\dot{\Delta}_h-\dot{\Delta}_l) + (\nu_h-\nu_l) \\ - \lambda_p\left[\int_{\underline{c}}^{\Delta_h}(\Delta_h-c)g(c)dc - \int_{\underline{c}}^{\Delta_l}(\Delta_l-c)g(c)dc\right] \\ - \lambda_s\Big[(1-\phi_l)(1-\theta) + \phi_h\theta\Big](\Delta_h-\Delta_l)\end{aligned}$$

This equation is analogous to the one we had before, and, in the same way, implies $\Delta_h > \Delta_l$, a contradiction of our assumption that $\Delta_l < \Delta_h$. □

A.2 Proof of Proposition 1

The general equilibrium effects of changing θ complicate the proof. We deal with it by first arguing that, in the sellers economy, the reservation value of high-type investors is independent of the other equilibrium objects, and in the buyers economy, the reservation value of low-type investors is independent of the other equilibrium objects. That is, in these two-limit cases we eliminate one of the four equilibrium equations, which reduces the general equilibrium effects we need to account for. Then we use the remaining three equations to sign the effects of moving θ between θ^A, θ^B, and θ^C in the equilibrium outcomes.

The equilibrium equations in steady state are

$$0 = (r+\mu)\Delta_l + \lambda_p \int_{\underline{c}}^{\Delta_l} (\Delta_l - c) g(c) dc - \lambda_s (1-\phi_h)(1-\theta)(\Delta_h - \Delta_l) - \nu_l \tag{26}$$

$$0 = \lambda_p (1-\phi_l) G(\Delta_l) - \mu \phi_l - \lambda_s \phi_l (1-\phi_h) \tag{27}$$

$$0 = (r+\mu)\Delta_h + \lambda_p \int_{\underline{c}}^{\Delta_h} (\Delta_h - c) g(c) dc + \lambda_s \phi_l \theta (\Delta_h - \Delta_l) - \nu_h \tag{28}$$

$$0 = \lambda_p (1-\phi_h) G(\Delta_h) - \mu \phi_h + \lambda_s \phi_l (1-\phi_h) \tag{29}$$

Proof of the buyer's economy *vs* the interior economy

$\boldsymbol{\Delta_l^C = \frac{\nu_l}{\mu+r}}$**:** The equilibrium equation (26) implies

$$(\mu+r)\Delta_l^C + \lambda_p \int_{\underline{c}}^{\Delta_l^C} (\Delta_l^C - c) g(c) dc = \nu_l.$$

This implies that low-types' reservation value in the buyer's economy is $\Delta_l^C = \frac{\nu_l}{\mu+r}$. The term $(\mu+r)\Delta_l + \lambda_p \int_{\underline{c}}^{\Delta_l} (\Delta_l - c) g(c) dc$ is strictly increasing as a function of Δ_l, and it is exactly ν_l when $\Delta_l^C = \frac{\nu_l}{\mu+r}$ because $\frac{\nu_l}{\mu+r} = \underline{c}$ so $\lambda_p \int_{\underline{c}}^{\Delta_l^C} (\Delta_l^C - c) g(c) dc$ equals zero. Therefore the only solution for the equality is $\Delta_l^C = \frac{\nu_l}{\mu+r}$.

$\boldsymbol{\Delta_l^B > \Delta_l^C}$**:** Because of depreciation ($\mu > 0$), there is no steady state where $\phi_h = 1$, so $1 - \phi_h^B$ is strictly positive. From lemma 1 we must have $\Delta_h^B - \Delta_l^B$ strictly positive. Therefore, $\lambda_s (1-\phi_h^B)(1-\theta^B)(\Delta_h^B - \Delta_l^B) > 0$, and we can conclude that

$$(\mu+r)\Delta_l^C + \lambda_p \int_{\underline{c}}^{\Delta_l^C} (\Delta_l^C - c) g(c) dc < (r+\mu)\Delta_l^B + \lambda_p \int_{\underline{c}}^{\Delta_l^B} (\Delta_l^B - c) g(c) dc$$

from the equilibrium equation 26. Since the term $(\mu+r)\Delta_l + \lambda_p \int_{\underline{c}}^{\Delta_l}(\Delta_l - c)g(c)dc$ is strictly increasing as a function of Δ_l, the above inequality implies that $\Delta_l^B > \Delta_l^C$.

$\phi_l^C = 0$: From the equilibrium equation 27 we have

$$0 = \lambda_p(1-\phi_l^C)G(\Delta_l^C) - \mu\phi_l^C - \lambda_s\phi_l^C(1-\phi_h^C) = -\phi_l^C[\mu + \lambda_s(1-\phi_h^C)].$$

There is no issuance of assets to dealers since $G(\Delta_l^C) = G\left(\nu_l/(\mu+r)\right) = G(\underline{c}) = 0$, and $\mu + \lambda_s(1-\phi_h^C) > 0$ since ϕ_h^C is smaller than one due to depreciation. Therefore, ϕ_l^C is zero.

$\phi_l^B > \phi_l^C$: From the equilibrium equation 27 we have

$$0 = \lambda_p(1-\phi_l^B)G(\Delta_l^B) - \mu\phi_l^B - \lambda_s\phi_l^B(1-\phi_h^B) \implies \phi_l^B = \frac{\lambda_p G(\Delta_l^B)}{\lambda_p G(\Delta_l^B) + \mu + \lambda_s(1-\phi_h^B)}.$$

We know that $G(\Delta_l^B) > 0$ since $\Delta_l^B > \Delta_l^C = \underline{c}$. Therefore, $\phi_l^B > 0 = \phi_l^C$.

$\Delta_h^C > \Delta_h^B$: We showed that $\phi_l^B > 0$, lemma 1 says that $\Delta_h^B - \Delta_l^B > 0$ and, therefore, $\lambda_s\phi_l^B\theta^B(\Delta_h^B - \Delta_l^B) > 0$. We showed that $\phi_l^C = 0$ so $\lambda_s\phi_l^C\theta^C(\Delta_h^C - \Delta_l^C) = 0$. The equilibrium equation 28, $\lambda_s\phi_l^B\theta^B(\Delta_h^B - \Delta_l^B) > 0$ and $\lambda_s\phi_l^C\theta^C(\Delta_h^C - \Delta_l^C) = 0$ imply

$$(r+\mu)\Delta_h^B + \lambda_p\int_{\underline{c}}^{\Delta_h^B}(\Delta_h^B - c)g(c)dc < (r+\mu)\Delta_h^C + \lambda_p\int_{\underline{c}}^{\Delta_h^C}(\Delta_h^C - c)g(c)dc.$$

Since the term $(r+\mu)\Delta_h + \lambda_p\int_{\underline{c}}^{\Delta_h}(\Delta_h - c)g(c)dc$ is strictly increasing as a function of Δ_h, the above inequality implies that $\Delta_h^B < \Delta_h^C$.

$\phi_h^B > \phi_h^C$: Define the function $F(\phi_l, \phi_h, \Delta_h; \Delta_l)$ as

$$F(\phi_l, \phi_h, \Delta_h; \Delta_l) = \begin{bmatrix} \lambda_p(1-\phi_l)G(\Delta_l) - \mu\phi_l - \lambda_s\phi_l(1-\phi_h) \\ \lambda_p(1-\phi_h)G(\Delta_h) - \mu\phi_h + \lambda_s\phi_l(1-\phi_h) \\ (r+\mu)\Delta_h + \lambda_p\int_{\underline{c}}^{\Delta_h}(\Delta_h - c)g(c)dc + \lambda_s\phi_l\theta^B(\Delta_h - \Delta_l) - \nu_h \end{bmatrix}. \tag{30}$$

Note that $F(\phi_l^B, \phi_h^B, \Delta_h^B; \Delta_l^B) = \mathbf{0}$ and $F(\phi_l^C, \phi_h^C, \Delta_h^C; \Delta_l^C) = \mathbf{0}$, where $\mathbf{0}$ is the zero column vector in $\mathbb{R}^3$. The first equality comes from the equilibrium definition, while the second comes from the equilibrium definition and $\phi_l^C = 0$. The equality $F(\phi_l, \phi_h, \Delta_h; \Delta_l) = \mathbf{0}$ implicitly defines ϕ_h as functions of Δ_l, and we can use the implicit function theorem to compute $\partial\phi_h/\partial\Delta_l$. Since $\Delta_l^B > \Delta_l^C$, if $\partial\phi_h/\partial\Delta_l$ is positive we can conclude that $\phi_h^B > \phi_h^C$.

To apply the implicit function theorem let us compute $D = det(\partial F/\partial(\phi_l, \phi_h, \Delta_h))$. We have

$$\frac{\partial F}{\partial(\phi_l, \phi_h, \Delta_h)} = \begin{bmatrix} -[\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] & \lambda_s \phi_l & 0 \\ \lambda_s(1-\phi_h) & -[\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] & \lambda_p(1-\phi_h) g(\Delta_h) \\ \lambda_s \theta^B (\Delta_h - \Delta_l) & 0 & r + \mu + \lambda_p G(\Delta_h) + \lambda_s \phi_l \theta^B \end{bmatrix},$$

and $D = det\left(\partial F/\partial(\phi_l, \phi_h, \Delta_h)\right)$ is

$$\begin{aligned}
D &= [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] \times [r + \mu + \lambda_p G(\Delta_h) + \lambda_s \phi_l \theta^B] \\
&\quad + \lambda_s \phi_l \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \theta^B (\Delta_h - \Delta_l) - \lambda_s \phi_l \times \lambda_s (1-\phi_h) \times [r + \mu + \lambda_p G(\Delta_h) + \lambda_s \phi_l \theta^B] \\
&= [r + \mu + \lambda_p G(\Delta_h) + \lambda_s \phi_l \theta^B] \Big\{ [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] \\
&\quad - \lambda_s \phi_l \lambda_s (1-\phi_h) \Big\} + \lambda_s \phi_l \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \theta^B (\Delta_h - \Delta_l) \\
&= [r + \mu + \lambda_p G(\Delta_h) + \lambda_s \phi_l \theta^B] \Big\{ [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\lambda_p G(\Delta_h) + \mu] + [\lambda_p G(\Delta_l) + \mu] \lambda_s \phi_l \\
&\quad + \cancel{\lambda_s (1-\phi_h) \lambda_s \phi_l} - \cancel{\lambda_s \phi_l \lambda_s (1-\phi_h)} \Big\} + \lambda_s \phi_l \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \theta^B (\Delta_h - \Delta_l) \\
&= [(r+\mu) + \lambda_p G(\Delta_h) + \lambda_s \phi_l \theta^B] \Big\{ [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\lambda_p G(\Delta_h) + \mu] \\
&\quad + [\lambda_p G(\Delta_l) + \mu] \lambda_s \phi_l \Big\} + \lambda_s \phi_l \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \theta^B (\Delta_h - \Delta_l) \geq (\mu + r)\mu^2 > 0.
\end{aligned}$$

Since $D \geq (\mu + r)\mu^2 > 0$ is bounded away from zero, we can apply the implicit function theorem all the way from Δ_l^C to Δ_l^B. The implicit function theorem implies that

$$\underbrace{\frac{\partial F}{\partial(\phi_l, \phi_h, \Delta_h)}}_{\text{matrix A}} \begin{bmatrix} \partial\phi_l/\partial\Delta_l \\ \partial\phi_h/\partial\Delta_l \\ \partial\Delta_h/\partial\Delta_l \end{bmatrix} = -\frac{\partial F}{\partial \Delta_l} = \underbrace{\begin{bmatrix} -\lambda_p(1-\phi_l) g(\Delta_l) \\ 0 \\ \lambda_s \phi_l \theta^B \end{bmatrix}}_{\text{vector b}}.$$

We can easily :| solve this system using Cramer's rule. We already computed $D = det(A)$. Let us compute D_{ϕ_h}, which is the determinant of the matrix A after replacing the second column of A with the vector b.

$$\begin{aligned}
D_{\phi_h} = &- \lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \theta^B (\Delta_h - \Delta_l) \\
&+ \lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_s (1-\phi_h) \times [r + \mu + \lambda_p G(\Delta_h) + \lambda_s \phi_l \theta^B] \\
&+ [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \phi_l \theta^B \\
= &- \lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \theta^B (-\Delta_l) \\
&+ \lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_s (1-\phi_h) \times [r + \mu + \lambda_p G(\Delta_h) - \lambda_p \theta^B g(\Delta_h) \Delta_h + \lambda_s \phi_l \theta^B] \\
&+ [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \phi_l \theta^B \\
= &\ \lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_p (1-\phi_h) g(\Delta_h) \times \lambda_s \theta^B \Delta_l
\end{aligned}$$

$$+\lambda_p(1-\phi_l)g(\Delta_l)\times\lambda_s(1-\phi_h)\times[r+\mu+\lambda_p(1-\theta^B)G(\Delta_h)+\lambda_s\phi_l\theta^B]$$
$$+[\lambda_pG(\Delta_l)+\mu+\lambda_s(1-\phi_h)]\times\lambda_p(1-\phi_h)g(\Delta_h)\times\lambda_s\phi_l\theta^B>0$$

From Cramer's rule $\partial\phi_h/\partial\Delta_l = {}^{D_{\phi_h}}/_{D} > 0$ and, therefore, $\phi_h^B > \phi_h^C$.

Proof of the seller's economy *vs* the interior economy

$\boldsymbol{\Delta_h^A > \Delta_h^B}$ **:** We showed that ϕ_l^B is strictly positive, and lemma 1 states that $\Delta_h^B - \Delta_l^B$ is strictly positive. These results imply that $\lambda_s\phi_l^B\theta^B(\Delta_h^B-\Delta_l^B)$ is strictly positive. The term $\lambda_s\phi_l^A\theta^A(\Delta_h^A-\Delta_l^A)$ is zero because θ^A is zero. From the inequality $\lambda_s\phi_l^B\theta^B(\Delta_h^B-\Delta_l^B)>0$, the equality $\lambda_s\phi_l^A\theta^A(\Delta_h^A-\Delta_l^A)=0$, and the equilibrium equation 28, we conclude that

$$(r+\mu)\Delta_h^B+\lambda_p\int_{\underline{c}}^{\Delta_h^B}(\Delta_h^B-c)g(c)dc<(r+\mu)\Delta_h^A+\lambda_p\int_{\underline{c}}^{\Delta_h^A}(\Delta_h^A-c)g(c)dc.$$

The term $(r+\mu)\Delta_h+\lambda_p\int_{\underline{c}}^{\Delta_h}(\Delta_h-c)g(c)dc$ is strictly increasing as a function of Δ_h. As a result, the above inequality implies that $\Delta_h^B<\Delta_h^A$.

$\boldsymbol{\phi_l^A > \phi_l^B}$**,** $\boldsymbol{\phi_h^A > \phi_h^B}$**, and** $\boldsymbol{\Delta_l^A > \Delta_l^B}$ **:** With abuse of notation, let us now define the function $F(\phi_l,\phi_h,\Delta_l;\theta,\Delta_h)$ as

$$F(\phi_l,\phi_h,\Delta_l;\theta,\Delta_h)=\begin{bmatrix}\lambda_p(1-\phi_l)G(\Delta_l)-\mu\phi_l-\lambda_s\phi_l(1-\phi_h)\\ \lambda_p(1-\phi_h)G(\Delta_h)-\mu\phi_h+\lambda_s\phi_l(1-\phi_h)\\ (r+\mu)\Delta_l+\lambda_p\int_{\underline{c}}^{\Delta_l}(\Delta_l-c)g(c)dc-\lambda_s(1-\phi_h)(1-\theta)(\Delta_h-\Delta_l)-\nu_l\end{bmatrix}. \quad (31)$$

It is easy to check that $F(\phi_l^A,\phi_h^A,\Delta_l^A;\theta^A,\Delta_h^A)=\mathbf{0}$ and $F(\phi_l^B,\phi_h^B,\Delta_l^B;\theta^B,\Delta_h^B)=\mathbf{0}$; the two equalities come from the equilibrium definition.

The equality $F(\phi_l,\phi_h,\Delta_l;\theta,\Delta_h)=\mathbf{0}$ implicitly defines ϕ_l, ϕ_h, and Δ_l as functions of θ and Δ_h. So we can use the implicit function theorem to compute $\partial\phi_l/\partial\Delta_h$, $\partial\phi_h/\partial\Delta_h$, $\partial\Delta_l/\partial\Delta_h$, $\partial\phi_l/\partial\theta$, $\partial\phi_h/\partial\theta$ and $\partial\Delta_l/\partial\theta$. We know that $\Delta_l^A>\Delta_l^B$ and $\theta^A<\theta^B$. Therefore, if $\partial\phi_l/\partial\Delta_h$, $\partial\phi_h/\partial\Delta_h$, and $\partial\Delta_l/\partial\Delta_h$ are positive, and $\partial\phi_l/\partial\theta$, $\partial\phi_h/\partial\theta$, and $\partial\Delta_l/\partial\theta$ are negative, we can conclude that $\phi_l^A>\phi_l^B$, $\phi_h^A>\phi_h^B$, and $\Delta_l^A>\Delta_l^B$.

To apply the implicit function theorem let us compute $D=det(\partial F/\partial(\phi_l,\phi_h,\Delta_l))$. We have

$$\frac{\partial F}{\partial(\phi_l,\phi_h,\Delta_l)}=\begin{bmatrix}-[\lambda_pG(\Delta_l)+\mu+\lambda_s(1-\phi_h)] & \lambda_s\phi_l & \lambda_p(1-\phi_l)g(\Delta_l)\\ \lambda_s(1-\phi_h) & -[\lambda_pG(\Delta_h)+\mu+\lambda_s\phi_l] & 0\\ 0 & \lambda_s(1-\theta)(\Delta_h-\Delta_l) & r+\mu+\lambda_pG(\Delta_l)+\lambda_s(1-\phi_h)(1-\theta)\end{bmatrix},$$

and $D = det\left(\partial F/\partial(\phi_l,\phi_h,\Delta_l)\right)$ is

$$\begin{aligned}
D =& [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \\
&+ \lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_s(1-\phi_h) \times \lambda_s(1-\theta)(\Delta_h - \Delta_l) \\
&- \lambda_s \phi_l \times \lambda_s(1-\phi_h) \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \\
=& [\lambda_p G(\Delta_l) + \mu] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \\
&+ \lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_s(1-\phi_h) \times \lambda_s(1-\theta)(\Delta_h - \Delta_l) \\
&- \cancel{\lambda_s \phi_l \times \lambda_s(1-\phi_h)} \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \\
&+ \cancel{\lambda_s(1-\phi_h) \times \lambda_s \phi_l} \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \\
&+ \lambda_s(1-\phi_h) \times [\lambda_p G(\Delta_h) + \mu] \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \\
=& [\lambda_p G(\Delta_l) + \mu] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \\
&+ \lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_s(1-\phi_h) \times \lambda_s(1-\theta)(\Delta_h - \Delta_l) \\
&+ \lambda_s(1-\phi_h) \times [\lambda_p G(\Delta_h) + \mu] \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \geq \mu^2(r+\mu) > 0
\end{aligned}$$

Since $D \geq (\mu + r)\mu^2 > 0$ is bounded away from zero we can apply the implicit function theorem all the way from Δ_h^A to Δ_h^B and θ^A to θ^B.

The implicit function theorem implies that

$$\underbrace{\frac{\partial F}{\partial(\phi_l,\phi_h,\Delta_l)}}_{\text{matrix A}} \begin{bmatrix} \partial\phi_l/\partial\Delta_h \\ \partial\phi_h/\partial\Delta_h \\ \partial\Delta_l/\partial\Delta_h \end{bmatrix} = -\frac{\partial F}{\partial \Delta_h} = \underbrace{\begin{bmatrix} 0 \\ -\lambda_p(1-\phi_h) g(\Delta_h) \\ \lambda_s(1-\phi_h)(1-\theta) \end{bmatrix}}_{\text{vector } b_h}.$$

and

$$\underbrace{\frac{\partial F}{\partial(\phi_l,\phi_h,\Delta_l)}}_{\text{matrix A}} \begin{bmatrix} \partial\phi_l/\partial\theta \\ \partial\phi_h/\partial\theta \\ \partial\Delta_l/\partial\theta \end{bmatrix} = -\frac{\partial F}{\partial \theta} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ -\lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \end{bmatrix}}_{\text{vector } b_\theta}.$$

We can solve the systems using Cramer's rule.

- Label $D_{\phi_l}^{\Delta_h}$ the determinant of A after replacing the first column of A with b_h.

$$\begin{aligned}
D_{\phi_l}^{\Delta_h} =& -\lambda_p(1-\phi_l) g(\Delta_l) \times \lambda_p(1-\phi_h) g(\Delta_h) \times \lambda_s(1-\theta)(\Delta_h - \Delta_l) \\
&+ \lambda_p(1-\phi_l) g(\Delta_l) \times [\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] \times \lambda_s(1-\phi_h)(1-\theta) > 0
\end{aligned}$$

From Cramer's rule $\partial\phi_l/\partial\Delta_h = D_{\phi_l}^{\Delta_h}/D > 0$. Label $D_{\phi_l}^{\theta}$ the determinant of A after replacing

the first column of A with b_θ.

$$D^{\theta}_{\phi_l} = -\lambda_p(1-\phi_l)g(\Delta_l) \times [\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) < 0$$

From Cramer's rule $\partial\phi_l/\partial\theta = D^{\theta}_{\phi_l}/D < 0$. Since $\partial\phi_l/\partial\Delta_h$ is positive and $\partial\phi_l/\partial\theta$ is negative, we can conclude that $\phi_l^A > \phi_l^B$.

- Label $D^{\Delta_h}_{\phi_h}$ the determinant of A after replacing the second column of A with b_h.

$$\begin{aligned} D^{\Delta_h}_{\phi_h} = {} & [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p(1-\phi_h)g(\Delta_h) \times [r + \mu + \lambda_p G(\Delta_l) + \lambda_s(1-\phi_h)(1-\theta)] \\ & + \lambda_p(1-\phi_l)g(\Delta_l) \times \lambda_s(1-\phi_h) \times \lambda_s(1-\phi_h)(1-\theta) > 0 \end{aligned}$$

From Cramer's rule $\partial\phi_h/\partial\Delta_h = D^{\Delta_h}_{\phi_h}/D > 0$. Label $D^{\theta}_{\phi_h}$ the determinant of A after replacing the second column of A with b_θ.

$$D^{\theta}_{\phi_l} = -\lambda_p(1-\phi_l)g(\Delta_l) \times \lambda_s(1-\phi_h) \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) < 0$$

From Cramer's rule $\partial\phi_h/\partial\theta = D^{\theta}_{\phi_l}/D < 0$. Since $\partial\phi_h/\partial\Delta_h$ is positive and $\partial\phi_h/\partial\theta$ is negative, we can conclude that $\phi_h^A > \phi_h^B$.

- Label $D^{\Delta_h}_{\Delta_l}$ the determinant of A after replacing the third column of A with b_h.

$$\begin{aligned} D^{\Delta_h}_{\Delta_l} = {} & [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s \phi_l] \times \lambda_s(1-\phi_h)(1-\theta) \\ & - \lambda_s \phi_l \times \lambda_s(1-\phi_h) \times \lambda_s(1-\phi_h)(1-\theta) \\ & - [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p(1-\phi_h)g(\Delta_h) \times \lambda_s(1-\theta)(\Delta_h - \Delta_l) \\ = {} & [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\mu + \lambda_s \phi_l] \times \lambda_s(1-\phi_h)(1-\theta) \\ & - \lambda_s \phi_l \times \lambda_s(1-\phi_h) \times \lambda_s(1-\phi_h)(1-\theta) \\ & + [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p G(\Delta_h) \times \lambda_s(1-\phi_h)(1-\theta) \\ & - [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p(1-\phi_h) \times \lambda_s(1-\theta)g(\Delta_h)\Delta_h \\ & + [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p(1-\phi_h)g(\Delta_h) \times \lambda_s(1-\theta)\Delta_l \\ = {} & [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\mu + \lambda_s \phi_l] \times \lambda_s(1-\phi_h)(1-\theta) \\ & - \lambda_s \phi_l \times \lambda_s(1-\phi_h) \times \lambda_s(1-\phi_h)(1-\theta) \\ & + [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \cancel{\lambda_p \lambda_s(1-\theta)(1-\phi_h) \times G(\Delta_h)} \\ & - [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \cancel{\lambda_p \lambda_s(1-\theta)(1-\phi_h) \times G(\Delta_h)} \\ & + [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p(1-\phi_h)g(\Delta_h) \times \lambda_s(1-\theta)\Delta_l \\ = {} & [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\mu + \lambda_s \phi_l] \times \lambda_s(1-\phi_h)(1-\theta) \end{aligned}$$

$$
\begin{aligned}
&- \lambda_s\phi_l \times \lambda_s(1-\phi_h) \times \lambda_s(1-\phi_h)(1-\theta) \\
&+ [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p(1-\phi_h)g(\Delta_h) \times \lambda_s(1-\theta)\Delta_l \\
=\ & [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \mu \times \lambda_s(1-\phi_h)(1-\theta) \\
&+ [\lambda_p G(\Delta_l) + \mu] \times \lambda_s\phi_l \times \lambda_s(1-\phi_h)(1-\theta) \\
&+ \lambda_s\phi_l \times \cancel{\lambda_s(1-\phi_h)} \times \lambda_s(1-\phi_h)(1-\theta) \\
&- \lambda_s\phi_l \times \cancel{\lambda_s(1-\phi_h)} \times \lambda_s(1-\phi_h)(1-\theta) \\
&+ [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p(1-\phi_h)g(\Delta_h) \times \lambda_s(1-\theta)\Delta_l \\
=\ & [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \mu \times \lambda_s(1-\phi_h)(1-\theta) \\
&+ [\lambda_p G(\Delta_l) + \mu] \times \lambda_s\phi_l \times \lambda_s(1-\phi_h)(1-\theta) \\
&+ [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times \lambda_p(1-\phi_h)g(\Delta_h) \times \lambda_s(1-\theta)\Delta_l > 0
\end{aligned}
$$

From Cramer's rule $\partial\Delta_l/\partial\Delta_h = D^{\Delta_h}_{\Delta_l}/D > 0$. Label $D^{\theta}_{\Delta_l}$ the determinant of the matrix A after replacing the third column of A with the vector b_θ.

$$
\begin{aligned}
D^{\theta}_{\Delta_l} =\ & - [\lambda_p G(\Delta_l) + \mu + \lambda_s(1-\phi_h)] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s\phi_l] \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \\
&+ \lambda_s\phi_l \times \lambda_s(1-\phi_h) \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \\
=\ & - [\lambda_p G(\Delta_l) + \mu] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s\phi_l] \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \\
&- \lambda_s(1-\phi_h) \times [\lambda_p G(\Delta_h) + \mu + \lambda_s\phi_l] \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \\
&+ \lambda_s\phi_l \times \lambda_s(1-\phi_h) \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \\
=\ & - [\lambda_p G(\Delta_l) + \mu] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s\phi_l] \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \\
&- \lambda_s(1-\phi_h) \times [\lambda_p G(\Delta_h) + \mu] \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \\
&- \lambda_s(1-\phi_h) \times \lambda_s\phi_l \times \cancel{\lambda_s(1-\phi_h)(\Delta_h - \Delta_l)} \\
&+ \lambda_s\phi_l \times \lambda_s(1-\phi_h) \times \cancel{\lambda_s(1-\phi_h)(\Delta_h - \Delta_l)} \\
=\ & - [\lambda_p G(\Delta_l) + \mu] \times [\lambda_p G(\Delta_h) + \mu + \lambda_s\phi_l] \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) \\
&- \lambda_s(1-\phi_h) \times [\lambda_p G(\Delta_h) + \mu] \times \lambda_s(1-\phi_h)(\Delta_h - \Delta_l) < 0
\end{aligned}
$$

From Cramer's rule $\partial\Delta_l/\partial\theta = D^{\theta}_{\Delta_l}/D < 0$. Since $\partial\Delta_l/\partial\Delta_h$ is positive and $\partial\Delta_l/\partial\theta$ is negative, we can conclude that $\Delta_l^A > \Delta_l^B$.

A.3 Proof of Lemma 2

The Hamiltonian of the planner's problem is

$$
\begin{aligned}
H = \phi_l \nu_l + \phi_h \nu_h - \lambda_p \left[(\pi_l - \phi_l) \int_0^{c_l} c g(c) dc + (\pi_h - \phi_h) \int_0^{c_h} c g(c) dc \right] \\
+ \gamma_l \Big\{ \lambda_p G(c_l)(\pi_l - \phi_l) + \lambda_s \phi_h p_{hl} (\pi_l - \phi_l) - \mu \phi_l - \lambda_s (\pi_h - \phi_h) p_{lh} \phi_l \Big\} \\
+ \gamma_h \Big\{ \lambda_p G(c_h)(\pi_h - \phi_h) + \lambda_s \phi_l p_{lh} (\pi_h - \phi_h) - \mu \phi_h - \lambda_s (\pi_l - \phi_l) p_{hl} \phi_h \Big\},
\end{aligned}
$$

where γ_l and γ_h are the co-state variables associated with the constraints on the law of motion of the distribution of investors (1) and (2). First, consider the first-order conditions with respect to c_l and c_h;

$$
\begin{aligned}
0 = \frac{\partial H}{\partial c_l} = & -\lambda_p (\pi_l - \phi_l) g(c_l) c_l + \lambda_p (\pi_l - \phi_l) g(c_l) \gamma_l \quad \text{and} \\
0 = \frac{\partial H}{\partial c_h} = & -\lambda_p (\pi_h - \phi_h) g(c_h) c_h + \lambda_p (\pi_h - \phi_h) g(c_h) \gamma_h .
\end{aligned}
$$

It is immediate from these conditions that $\gamma_l = c_l$ and $\gamma_h = c_h$. The planner equalizes the marginal cost of issuing assets to either low or high types to the shadow prices or marginal social gains associated with the Lagrange multipliers of the feasibility constraints (1)-(2). The terms coinciding with the mass of agents affected cancel out in both the gain and cost. Now, consider the first-order conditions with respect to ϕ_l and ϕ_h;

$$
\begin{aligned}
r\gamma_l = \dot{\gamma}_l + \frac{\partial H}{\partial \phi_l} = & \dot{\gamma}_l + \nu_l - \mu \gamma_l \\
& - \lambda_p \int_0^{\gamma_l} (\gamma_l - c) g(c) dc + (\gamma_h - \gamma_l) \lambda_s \left[\phi_h p_{hl} + (\pi_h - \phi_h) p_{lh} \right], \\
r\gamma_h = \dot{\gamma}_h + \frac{\partial H}{\partial \phi_h} = & \dot{\gamma}_h + \nu_h - \mu \gamma_h \\
& - \lambda_p \int_0^{\gamma_h} (\gamma_h - c) g(c) dc - (\gamma_h - \gamma_l) \lambda_s \left[\phi_l p_{lh} + (\pi_l - \phi_l) p_{hl} \right],
\end{aligned}
$$

where we have used the fact that $c_i = \gamma_i$ for $i \in \{l, h\}$. The above two equations imply that $\gamma_l < \gamma_h$. To see that this is the case, note that we can write the difference $\gamma_h - \gamma_l$ as

$$
\begin{aligned}
r(\gamma_h - \gamma_l) = & (\dot{\gamma}_h - \dot{\gamma}_l) + \nu_h - \nu_l - \mu(\gamma_h - \gamma_l) \\
& - \lambda_p \left[\int_0^{\gamma_h} (\gamma_h - c) g(c) dc - \int_0^{\gamma_l} (\gamma_l - c) g(c) dc \right] \\
& - (\gamma_h - \gamma_l) \lambda_s \left[(\pi_l - \phi_l) p_{hl} + \phi_l p_{lh} \right] - (\gamma_h - \gamma_l) \lambda_s \left[\phi_h p_{hl} + (\pi_h - \phi_h) p_{lh} \right].
\end{aligned}
$$

Suppose to the contrary that $\gamma_l \geq \gamma_h$. Then the left hand side of the above equation would be smaller or equal to zero. However, in order to have the right hand side smaller or equal to zero, $\dot{\gamma}_h - \dot{\gamma}_l$ would have to be strictly negative. Since $\gamma_h \geq 0$, this would imply that γ_l needs to converge to infinity at a rate higher than r, which would imply an explosive path for γ_l (that is, $\lim e^{-rt}\gamma_l = \infty$), a violation of the transversality condition. Hence, $\gamma_l < \gamma_h$. Finally, consider the first-order conditions with respect to the trade probabilities, p_{hl} and p_{lh};

$$\begin{aligned}\frac{\partial H}{\partial p_{lh}} &= -\gamma_l \lambda_s \phi_l(\pi_h - \phi_h) + \gamma_h \lambda_s \phi_l(\pi_h - \phi_h) \geq 0 \\ \frac{\partial H}{\partial p_{hl}} &= \gamma_l \lambda_s(\pi_l - \phi_l)\phi_h - \gamma_h \lambda_s(\pi_l - \phi_l)\phi_h \leq 0.\end{aligned}$$

with complementary slackness with multipliers of constraints that $p_{ij} \in [0,1]$ for $i,j = l,h$. Notice, since $\gamma_h > \gamma_l$, that the inequalities must hold with a strict inequality implying that $p_{lh} = 1$ and $p_{hl} = 0$. In words, it is never optimal for the planner to reallocate assets from high-type investors to low-type investors since the shadow value of high types is always strictly larger than that of low types.

A.4 Proof of Proposition 2

To see this is the case, let an asset allocation $\{\boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{p}\}$ combined with reservation values Δ_l and Δ_h be a decentralized equilibrium and assume, by way of contradiction, that this asset allocation solves the planner's problem. By Proposition 2, there exist co-state variables γ_l and γ_h such that γ_l, γ_h, ϕ_l, and ϕ_h solve the system of differential equations (14)-(17). By Definition 1, Δ_l, Δ_h, ϕ_l, and ϕ_h solve the system of differential equations (9)-(12). Equations (15) and (10) imply that

$$\begin{aligned}\dot{\phi}_l &= \lambda_p(\pi_l - \phi_l)G(\gamma_l) - \mu\phi_l - \lambda_s\phi_l(\pi_h - \phi_h) && (32)\\ &= \lambda_p(\pi_l - \phi_l)G(\Delta_l) - \mu\phi_l - \lambda_s\phi_l(\pi_h - \phi_h), && (33)\end{aligned}$$

or $\gamma_l = \Delta_l$. Analogously, equations (17) and (12) imply that

$$\begin{aligned}\dot{\phi}_h &= \lambda_p(\pi_h - \phi_h)G(\gamma_h) - \mu\phi_h + \lambda_s\phi_l(\pi_h - \phi_h) && (34)\\ &= \lambda_p(\pi_h - \phi_h)G(\Delta_h) - \mu\phi_h + \lambda_s\phi_l(\pi_h - \phi_h), && (35)\end{aligned}$$

or $\gamma_h = \Delta_h$. The above two results, together with (14)-(16) and (9)-(11), imply that

$$\begin{aligned}
r\gamma_l &= \dot{\gamma}_l + \nu_l - \mu\gamma_l - \lambda_p \int_0^{\Delta_l} (\Delta_l - c) g(c) dc + \lambda_s (\pi_h - \phi_h)(\gamma_h - \gamma_l) \\
&= \dot{\gamma}_l + \nu_l - \mu\gamma_l - \lambda_p \theta_a \int_0^{\Delta_l} (\Delta_l - c) g(c) dc + \lambda_s (\pi_h - \phi_h)(1-\theta)(\gamma_h - \gamma_l), \quad \text{and} \\
r\gamma_h &= \dot{\gamma}_h + \nu_h - \mu\gamma_h - \lambda_p \int_0^{\Delta_h} (\Delta_h - c) g(c) dc - \lambda_s \phi_l (\gamma_h - \gamma_l) \\
&= \dot{\gamma}_h + \nu_h - \mu\gamma_h - \lambda_p \theta_a \int_0^{\Delta_h} (\Delta_h - c) g(c) dc - \lambda_s \phi_l \theta (\gamma_h - \gamma_l).
\end{aligned}$$

If our candidate asset allocation coincides with a decentralized equilibrium, then we must be able to find constants θ_a and θ that solve (32)-(34). We can write the system as

$$\begin{aligned}
&\lambda_p \int_0^{\Delta_l} (\Delta_l - c) g(c) dc \theta_a + \lambda_s (\pi_h - \phi_h)(\gamma_h - \gamma_l)\theta = \lambda_p \int_0^{\Delta_l} (\Delta_l - c) g(c) dc \quad \text{and} \\
&\lambda_p \theta_a \int_0^{\Delta_h} (\Delta_h - c) g(c) dc \theta_a + \lambda_s \phi_l (\gamma_h - \gamma_l)\theta = \lambda_p \int_0^{\Delta_h} (\Delta_h - c) g(c) dc + \lambda_s \phi_l (\gamma_h - \gamma_l).
\end{aligned}$$

Additionally, since θ_a and θ are bargaining powers, we must have that $\theta_a, \theta \in [0,1]$. Moreover, it is easy to show that a solution of the differential equations discussed above cannot have either $\gamma_l = 0$, $\gamma_h = 0$, $\phi_h = \pi_h$, or $\phi_l = 0$ for all but a measure zero of period t's. Therefore, $\lambda_p \int_0^{\Delta_h} (\Delta_h - c) g(c) dc$ and $\lambda_s \phi_l (\gamma_h - \gamma_l)$ most both be strictly positive and the only $\theta_a, \theta \in [0,1]$ that satisfy the second equation in the above system are $\theta_a = \theta = 1$. But note that $\theta_a = \theta = 1$ does not solve the first equation in this system. Which contradicts that the asset allocation solves the planner's problem.

A.5 Proof of Proposition 3

The proof of proposition 3 is analogous to the proof of proposition 1, and we omit it here.

A.6 Proof of Proposition 4

We know by Lemma 2 that $\{\phi_l^*, \phi_h^*, \gamma_l, \gamma_h\}$ solves the differential equations (14)-(17). Then, after replacing $\boldsymbol{\tau}$, we can see that $\{\phi_l^*, \phi_h^*, \gamma_l, \gamma_h\}$ also solves the decentralized equilibrium with intervention equations (19)-(22). And $\boldsymbol{\tau}$ is feasible because $2\bar{\tau} = \phi_l \tau_l - \phi_h \tau_h$.

B Appendix - Municipal Bond Market Data

This section describes the moments we use in Section 6. We target five moments from the US municipal bond market: average maturity, average yield, the relative size of the secondary market to the primary market, the average bid-ask spread, and the average time between buying an asset in the primary market and selling in the secondary market.

The first three moments are computed using data available from the Municipal Securities Rulemaking Board (MSRB). The MSRB requires securities dealers, issuers, and those acting on their behalf to submit information on municipal bond trades and disclosure documents for all transactions of municipal bonds within 15 minutes of the time of trade. The data are publicly available through the MSRB's Electronic Municipal Market Access (EMMA) portal, however we obtain historical data through Wharton Research Data Services. Our dataset includes transaction-level information for all trades of municipal bonds involving a securities dealer in which the municipal bond was assigned a unique CUSIP identification number. This nearly covers the universe of municipal bond trades. Our sample includes transactions from January 3, 2005, through December 31, 2014. We drop transactions that are missing a par value or price. We also drop all variable rate securities since we do not see any information about the current interest rate at the time of the transaction. In our data, 70% of all transactions are fixed rate or discount bonds (0-coupon bonds). We also drop transactions whose par value is less than $1,000. In this sample selection, we are left with 86.4 million observations on 1.89 million unique bond issues where we identify a bond issue by its unique CUSIP identifier.[12]

The average maturity of the bond is directly given in the data. Following Green et al. (2007), we identify transactions occurring in the first 90 days after the dated date as primary market trades. Doing so implies that 62% of all trades by volume occur in the primary market and 38% occur in the secondary market. Finally, we calculate the average yield using the following method. Seventy percent of transactions in the data report the yield-to-worst, defined as the lowest of the yield calculated to the call option, par option, or maturity. The relevant yield in terms of our model is the yield-to-expected-redemption, or a weighted average of the yield-to-call, yield-to-par, and yield-to-maturity, weighting by the respective probabilities of each event occurring. Since we do not observe this yield in the data, our procedure for calculating the yield is as follows. If the yield-to-worst is reported, we take that to be the yield generated in the model. When no yield is reported, we calculate a yield-to-maturity. For discount bonds, the yield-to-maturity, i, is given by

[12] New issuance of municipal debt is typically done through a series of bonds with differing maturity dates and coupon rates. Each bond within the series is given a unique CUSIP identification number, which serves as our definition of an issue.

the formula $price = par/(1+i)T$, where T is the time until maturity. For other fixed-rate bonds, we use the console formula $i = c/(price?c)$, where c is the coupon rate.

The last two moments are taken directly from Green et al. (2007). Using a similar MSRB dataset (although from May 2000 to January 2004) they estimate the average time municipal securities dealers hold inventories (time to sell) and the respective markup. Since the identity of dealers are not available in the MSRB data, they use several procedures to link buy and sell transactions. We use moments from their "full sample" that is the most inclusive.[13] The average gross markup of 1.69% is reported in Table 4, and the time it takes a dealer to intermediate an asset is approximately five days (see Table 5).

C Supplementary Appendix - not for publication

C.1 A model with a continuum of investor types

In this section we augment the economy to allow for a continuum of investor types, as studied in Hugonnier et al. (2014). Unlike the simple model with two types of investors, here all investors buy and sell assets in the secondary market, thus all of them acting as intermediaries. Because investors are in both sides of the market, all of them fail to internalize the full gains from trade when buying and selling, adding a new layer of inefficiency with respect to the simple model with two types where each type of investor failed to internalize the full gains from trade of either selling or buying assets. Still, we show that the decentralized equilibrium is always inefficient, provided that the gains from a particular trade meeting in the secondary market have to be fully split by the meeting participants. We also study an extension of the model where we allow the bargaining power to be fully dependent on the identities of trade participants in a given meeting, and we show that the decentralized equilibrium remains inefficient.

We call an investor not holding an asset a non-owner investor (subscript n in the equations below), and we call an investor holding an asset an owner investor (subscript o in the equations below). There is a measure two of investors, and let $F(\nu)$ denote the cumulative distribution of investor types, with support $[\underline{\nu}, \bar{\nu}]$. Likewise, we use $\Phi(\nu)$ to denote the cumulative distribution of owner investors, with $\Phi(\nu) \leq F(\nu)$ for all ν, with $\phi(\nu) \equiv \partial\Phi(\nu)/\partial\nu$. Further, let $p^a(c, \nu_n)$ denote the probability that an issuer with issuance cost c trades with a non owner investor of type ν_n at the primary market, and let $p^b(\nu_o, \nu_n)$ denote the probability that an owner investor of type ν_o trades with a non owner investor of type ν_n at the secondary market.

[13] For more information about their identification strategy, please see Green et al. (2007).

The evolution of the distribution $\Phi(\nu)$ is given by

$$\begin{aligned}\dot{\Phi}(\nu) = & -\mu\Phi(\nu) - \lambda_s \int_{\underline{\nu}}^{\nu} \int_{\nu}^{\bar{\nu}} p^b(\nu_o, \nu_n)[f(\nu_n) - \phi(\nu_n)]\phi(\nu_o) d\nu_o d\nu_n \\ & + \lambda_s \int_{\nu}^{\bar{\nu}} \int_{\underline{\nu}}^{\nu} p^b(\nu_o, \nu_n)[f(\nu_n) - \phi(\nu_n)]\phi(\nu_o) d\nu_o d\nu_n \\ & + \lambda_p \int_{\underline{c}}^{\bar{c}} p^a(c, \nu_n)[f(\nu_n) - \phi(\nu_n)]g(c) dc d\nu_n \ .\end{aligned}$$

The equation states that the fraction of owner investors of type less or equal to ν holding assets suffers an outflow in two ways. First, assets can mature. Second, owner investors with type $\nu_o \leq \nu$ can sell it to a non owner investor satisfying $\nu_n > \nu$. Likewise, the fraction of owner investors of type less or equal to ν holding assets gets an inflow also in two ways. First, a non owner investor of type $\nu_n \leq \nu$ can buy the asset in the primary market from an issuer. Second, a non owner investor of type $\nu_n \leq \nu$ buys the asset from an owner investor of type $\nu_o > \nu$. Notice that this last case will not be a feature of the decentralized equilibrium nor of the efficient allocation. It proves useful to provide the evolution of the density function $\phi(\nu)$,

$$\begin{aligned}\dot{\phi}(\nu) = & -\mu\phi(\nu) - \lambda_s \int_{\nu}^{\bar{\nu}} [f(\nu_n) - \phi(\nu_n)]\phi(\nu) d\nu_n + \lambda_s \int_{\underline{\nu}}^{\nu} p^b(\nu_o, \nu)[f(\nu) - \phi(\nu)]\phi(\nu_o) d\nu_o \\ & - \lambda_s \int_{\underline{\nu}}^{\nu} p^b(\nu, \nu_n)[f(\nu_n) - \phi(\nu_n)]\phi(\nu) d\nu_n + \lambda_s \int_{\nu}^{\bar{\nu}} p^b(\nu_o, \nu)[f(\nu) - \phi(\nu)]\phi(\nu_o) d\nu_o \\ & + \lambda_p \int_{\underline{c}}^{\bar{c}} p^a(c, \nu)[f(\nu) - \phi(\nu)]g(c) dc \ . \end{aligned} \tag{36}$$

We now solve for the decentralized equilibrium. We solve for the decentralized equilibrium under the following two assumptions: (i) $p^a(c, \nu_n) = 1$ if $c \leq \Delta(\nu_n)$ and $p^a(c, \nu_n) = 0$ otherwise, (ii) $p^b(\nu_o, \nu_n) = 1$ if $\nu_o \leq \nu_n$ and $p^b(\nu_o, \nu_n) = 0$ otherwise. Later we verify that these two assumptions are satisfied in the decentralized equilibrium. Under these assumptions, the equilibrium exhibits a simple pattern of trade in the both primary and secondary markets. In the primary market, issuers issue whenever they find a non owner investor with reservation value $\Delta(\nu_n)$ above their cost of issuance c. In the secondary market, owner investors of of type ν_n sell to any non owner investor they encounter in the secondary market, as long as $\nu_o > \nu_n$. These trade patterns are the analogous ones to those obtained in the simple model with only two types of investors. Using this observation, the

value functions for a non owner and owner investor of type ν are given by

$$rV_n(\nu) = \dot{V}_n(\nu) + \lambda_p \int_{\underline{c}}^{\Delta(\nu)} [\Delta(\nu) - c] g(c) dc + \lambda_s \theta \int_{\underline{\nu}}^{\nu} [\Delta(\nu) - \Delta(\nu_o)] \phi(\nu_o) d\nu_o \ ,$$

$$rV_o(\nu) = \dot{V}_o(\nu) + \nu - \mu\Delta(\nu) + \lambda_s(1-\theta) \int_{\nu}^{\bar{\nu}} [\Delta(\nu_n) - \Delta(\nu)] \{f(\nu_n) - \phi(\nu_n)\} d\nu_n \ ,$$

where $\Delta(\nu) \equiv V_o(\nu) - V_n(\nu)$. Using these expressions we can derive an expression for $\Delta(\nu)$,

$$\begin{aligned}(r+\mu)\Delta(\nu) = \dot{\Delta}(\nu) + \nu - \lambda_p \int_{\underline{c}}^{\Delta(\nu)} [\Delta(\nu) - c] g(c) dc \\ + \lambda_s(1-\theta) \int_{\nu}^{\bar{\nu}} [\Delta(\nu_n) - \Delta(\nu)] \{f(\nu_n) - \phi(\nu_n)\} d\nu_n \\ - \lambda_s \theta \int_{\underline{\nu}}^{\nu} [\Delta(\nu) - \Delta(\nu_o)] \phi(\nu_o) d\nu_o \ . \end{aligned} \tag{37}$$

It is easy to show that $\Delta(\nu)$ is increasing in ν, which validates the assumption made above regarding the trade patterns in the economy.[14] The expression for the reservation value is analogous to that one obtained for the model with two investor types, as presented in equations (9) and (11). The only difference is that while in the simple model low-valuation investors where sellers in the secondary market and high-valuation investors where buyers in the secondary market, in the model with continuum of types all investors are both buyers and sellers in the secondary market. Thus, all of them serve a role as intermediators, while that role was only assigned to the low-valuation investors in the model with two investor types. We next define a decentralized equilibrium in the economy with a continuum of investor types,

Definition 4. *A decentralized equilibrium is a set of trading protocols $p^a(c, \nu_n), p^b(\nu_o, \nu_n)$ for all $c \in [\underline{c}, \bar{c}]$, $\nu_n \in [\underline{\nu}, \bar{\nu}]$ and $\nu_o \in [\underline{\nu}, \bar{\nu}]$, an asset allocation and bounded reservation values for investors $\{\phi(v), \Delta(\nu)\}$ for all $\nu \in [\underline{\nu}, \bar{\nu}]$, that solve the system of differential equations given by equations (36) and (37), with initial conditions $\phi(\nu) = \phi^0(\nu)$, for all $\nu \in [\underline{\nu}, \bar{\nu}]$. The trading protocols satisfy (i) $p^a(c, \nu_n) = 1$ if $c \leq \Delta(\nu_n)$ and $p^a(c, \nu_n) = 0$ otherwise, and (ii) $p^b(\nu_o, \nu_n) = 1$ if $\nu_o \leq \nu_n$ and $p^b(\nu_o, \nu_n) = 0$ otherwise.*

We now solve for the efficient allocation. The efficient allocation solves the following problem,

$$\max_{\phi, p^a, p^b} \int_0^{\infty} e^{-rt} \left\{ \int_{\underline{\nu}}^{\bar{\nu}} \nu \phi(\nu) d\nu - \lambda_p \int_{\underline{\nu}}^{\bar{\nu}} \int_{\underline{c}}^{\bar{c}} p^a(c, \nu) c \{f(\nu) - \phi(\nu)\} g(c) dc d\nu \right\} dt \ ,$$

[14] The proof is analogous to the proof of lemma 1, and thus it is not provided.

subject to equation (36) for all $\nu \in [\underline{\nu}, \bar{\nu}]$ and $\int_{\underline{\nu}}^{\bar{\nu}} \phi(\nu) d\nu = 1$.

We solve for the efficient allocation by forming the Hamiltonian of the problem, $\mathscr{H}$. We use $\gamma(\nu)$ as the co-state variable for equation (36), and $\bar{\gamma}$ as the multiplier of the second restriction (i.e. the asset density must add up to one).

We begin by studying the optimal choices for the controls p^a and p^b. Differentiating the Hamiltonian with respect to $p^a(c, \nu)$ provides

$$\frac{\partial \mathscr{H}}{\partial p^a(c,\nu)} = \lambda_p [f(\nu) - \phi(\nu)]\{\gamma(\nu) - c\} g(c) dc d\nu \ .$$

This expression is positive if $\gamma(\nu) \geq c$ and negative otherwise. This provides that the efficient allocation requires that an issuer issues whenever he encounters a non owner investor with co-state variable $\gamma(\nu)$ above his cost c. That is, $p^a(c, \nu) = 1$ if $\gamma(\nu) \geq c$, and $p^a(c, \nu) = 0$ if $\gamma(\nu) < c$. Likewise, differentiating the Hamiltonian with respect to $p^b(\nu_0, \nu_n)$ provides

$$\frac{\partial \mathscr{H}}{\partial p^b(\nu_o,\nu_n)} = \lambda_s [f(\nu_n) - \phi(\nu_n)]\phi(\nu_o)\{\gamma(\nu_n) - \gamma(\nu_o)\} d\nu_o \nu_n \ ,$$

which is positive if $\gamma(\nu_n) \geq \gamma(\nu_o)$. This provides that the efficient allocation requires that an owner-investor of type ν_o sells whenever he encounters a non owner investor with co-state variable $\gamma(\nu_n)$ above his co-state variable $\gamma(\nu_n)$.

For the co-state variable $\gamma(\nu)$ we use that optimal control requires $\partial \mathscr{H} / \partial \gamma(\nu) = r\gamma(\nu) - \dot{\gamma}(\nu)$. After operating with this expression, we obtain the following expression,

$$\begin{aligned}(r+\mu)\gamma(\nu) =& \dot{\gamma}(\nu) + \nu - \lambda_p \int_{\underline{c}}^{\gamma(\nu)} [\gamma(\nu) - c] g(c) dc \\ &+ \lambda_s \int_{\nu}^{\bar{\nu}} [\gamma(\nu_n) - \gamma(\nu)]\{f(\nu_n) - \phi(\nu_n)\} d\nu_n \\ &- \lambda_s \int_{\underline{\nu}}^{\nu} [\gamma(\nu) - \gamma(\nu_o)]\phi(\nu_o) d\nu_o \ . \end{aligned} \tag{38}$$

The next definition describes an efficient allocation.

Definition 5. *An efficient allocation is a set of trading protocols $p^a(c, \nu_n), p^b(\nu_o, \nu_n)$ for all $c \in [\underline{c}, \bar{c}]$, $\nu_n \in [\underline{\nu}, \bar{\nu}]$, and $\nu_o \in [\underline{\nu}, \bar{\nu}]$, an asset allocation and co-state variables for investors $\{\phi(v), \gamma(\nu)\}$ for all $\nu \in [\underline{\nu}, \bar{\nu}]$, that solve the system of differential equations given by equations (36) and (38), with initial conditions $\phi(\nu) = \phi^0(\nu)$, for all $\nu \in [\underline{\nu}, \bar{\nu}]$. The trading protocols satisfy (i) $p^a(c, \nu_n) = 1$ if $c \leq \Delta(\nu_n)$ and $p^a(c, \nu_n) = 0$ otherwise, and (ii) $p^b(\nu_o, \nu_n) = 1$ if $\nu_o \leq \nu_n$ and $p^b(\nu_o, \nu_n) = 0$ otherwise.*

Because in a model with a continuum of types all investors resale assets, the inefficiency appears twice for each investor, rather than once as occurred in the model with two types. In the simpler model, low-valuation investors fail to internalize the full gains from trade when selling, while the high-valuation investors failed to internalize the full gains from trade when buying. In the model with a continuum of types, each type of investor fails to internalize both sources of gains from trade, as all of them buy and sell assets in the secondary market.

Lemma 3. *For any $\theta \in [0,1]$, the solution to equations (36) and (37) does not solve equations (36) and (38). That is, for any $\theta \in [0,1]$, the decentralized equilibrium is inefficient.*

Lemma 3 extends the inefficiency result obtained in the simple model to the model with a continuum of types. As in the simple model, the nature of the lack of efficiency stems from the failure of the decentralized equilibrium to internalize the full gains from trade when investors are buying and selling assets in the secondary market. As we show in the appendix, at a superficial level, there seems to be a choice of bargaining power that allows investors of type ν to fully internalize the gains from trade. This choice of bargaining power must be such that $\theta = \theta(\nu)$, so that different investor types have different bargaining powers. However, when a seller of type ν and buyer of type $\tilde{\nu} \geq \nu$ trade, their bargaining powers must add up to one, so that the surplus generated by the trade $\Delta(\tilde{\nu}) - \Delta(\nu)$ is fully divided by the trade participants. This restriction guarantees that it is not possible to choose a set of bargaining powers that depend on the investor type, so that the decentralized equilibrium is not efficient.

C.2 Trading partners

A natural conjecture that follows from the previous finding is that if the bargaining power were to be allowed to depend on the identity of both parties in a trade meeting, the decentralized equilibrium would be efficient for the appropriate choice of bargaining weights. With this in mind, in this section we augment the model to allow for the bargaining power to depend on the identity of trade participants. In particular, at the secondary market, for any seller of type ν_o and any buyer of type $\nu_n \geq \nu_o$ for all $\nu_o \in [\underline{\nu}, \bar{\nu}]$, let $\theta = \theta(\nu_o, \nu) \in [0,1]$ denote the bargaining power of the buyer in a meeting between these two investors.

In terms of the decentralized equilibrium, the differential equation for the reservation value presented in equation (37) is now given by

$$(r+\mu)\Delta(\nu) = \dot{\Delta}(\nu) + \nu - \lambda_p \int_{\underline{c}}^{\Delta(\nu)} [\Delta(\nu) - c] g(c) dc$$

$$+ \lambda_s \int_{\nu}^{\bar{\nu}} \{1 - \theta(\nu, \nu_n)\}[\Delta(\nu_n) - \Delta(\nu)]\{f(\nu_n) - \phi(\nu_n)\}d\nu_n$$
$$- \lambda_s \int_{\underline{\nu}}^{\nu} \theta(\nu_o, \nu)[\Delta(\nu) - \Delta(\nu_o)]\phi(\nu_o)d\nu_o \ . \tag{39}$$

All the other equations for the decentralized equilibrium remain unchanged.

Efficiency of the decentralized equilibrium requires that $\{\phi(\nu), \Delta(\nu)\}$ is also a solution to the efficient allocation problem. As before, this reduces to checking whether there is a way to choose $\theta(\nu_o, \nu_n)$ that makes equation (39) identical to (38). This reduces to

$$\int_{\nu}^{\bar{\nu}} \{1 - \theta(\nu, \nu_n)\}[\Delta(\nu_n) - \Delta(\nu)]\{f(\nu_n) - \phi(\nu_n)\}d\nu_n - \int_{\underline{\nu}}^{\nu} \theta(\nu_o, \nu)[\Delta(\nu) - \Delta(\nu_o)]\phi(\nu_o)d\nu_o$$
$$= \int_{\nu}^{\bar{\nu}} [\Delta(\nu_n) - \Delta(\nu)]\{f(\nu_n) - \phi(\nu_n)\}d\nu_n - \int_{\underline{\nu}}^{\nu} [\Delta(\nu) - \Delta(\nu_o)]\phi(\nu_o)d\nu_o \ .$$

This condition has to be satisfied for every $\nu \in [\underline{\nu}, \bar{\nu}]$. In particular, it has to be satisfied for $\nu = \underline{\nu}$ and $\nu = \bar{\nu}$. For investors with the lowest valuation $\nu = \underline{\nu}$, the condition reduces to

$$\int_{\underline{\nu}}^{\bar{\nu}} \theta(\underline{\nu}, \nu_n)[\Delta(\nu_n) - \Delta(\underline{\nu})]\{f(\nu_n) - \phi(\nu_n)\}d\nu_n = 0 \ .$$

Thus, $\theta(\underline{\nu}, \nu_n) = 0$ for all $\nu_n \in [\underline{\nu}, \bar{\nu}]$ as $[\Delta(\nu_n) - \Delta(\underline{\nu})]\{f(\nu_n) - \phi(\nu_n)\} > 0$ for all $\nu_n \in [\underline{\nu}, \bar{\nu}]$. That is, every trade in the secondary market that includes the lowest-valuation investor must assign all the gains from trade to the seller. This occurs because for an investor of type $\underline{\nu}$, trading in the secondary market only involves selling an asset, and thus the investor fails to internalize the full gains from trade when selling, which can only be corrected by giving investors of type $\underline{\nu}$ the full bargaining power.

For investors with the highest valuation $\nu = \bar{\nu}$, the previous condition is given by

$$\int_{\underline{\nu}}^{\bar{\nu}} \{1 - \theta(\nu_o, \nu)\}[\Delta(\bar{\nu}) - \Delta(\nu_o)]\phi(\nu_o)d\nu_o = 0 \ .$$

Using the same logic as before, this condition is only satisfied if $\theta(\nu_o, \bar{\nu}) = 1$ for all $\nu_o \in [\underline{\nu}, \bar{\nu}]$, so that whenever a trade includes the highest-valuation investors as buyers, the full gains from trade must go to the buyer. This is an immediate consequence of the fact that these investors only buy in the secondary market and thus fail to internalize the full gains from trade only when buying assets. This can only be corrected by giving investors of type $\bar{\nu}$ the full bargaining power.

A natural implication of these two-limit cases is that the decentralized equilibrium with trade-specific bargaining powers is also inefficient. Whenever an owner investor of type

$\underline{\nu}$ meets a non owner investor of type $\bar{\nu}$, efficiency would require that we provide full bargaining power to both buyer and seller, violating the restriction that investors can at most share the surplus generated in the trade at hand.