DISCUSSION PAPER SERIES

No. 5965

A SEARCH-BASED THEORY OF THE ON-THE-RUN PHENOMENON

Dimitri Vayanos and Pierre-Olivier Weill

FINANCIAL ECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP5965.asp**

ISSN 0265-8003

A SEARCH-BASED THEORY OF THE ON-THE-RUN PHENOMENON

Dimitri Vayanos, London School of Economics (LSE) and CEPR
Pierre-Olivier Weill, University of California, Los Angeles

Discussion Paper No. 5965
November 2006

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 5965

November 2006

ABSTRACT

A Search-Based Theory of the On-the-Run Phenomenon*

We propose a model in which assets with identical cash flows can trade at different prices. Infinitely-lived agents can establish long positions in a search spot market, or short positions by first borrowing an asset in a search repo market. We show that short-sellers can endogenously concentrate in one asset because of search externalities and the constraint that they must deliver the asset they borrowed. That asset enjoys greater liquidity, measured by search times, and a higher lending fee ('specialness'). Liquidity and specialness translate into price premia that are consistent with no-arbitrage. We derive closed-form solutions for small frictions, and can generate price differentials in line with observed on-the-run premia.

JEL Classification: D8 and G1
Keywords: asset pricing, liquidity, on-the-run bonds and search

Dimitri Vayanos
Department of Accounting and Finance
London School of Economics
Houghton Street
London
WC2A 2AE
Email: D.Vayanos@lse.ac.uk

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=116836

Pierre-Olivier Weill
Department of Economics
UCLA
Bunche Hall 8283
Box 951477
Los Angeles, CA 90095-1477
USA
Email: poweill@econ.ucla.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=162145

* We thank Tobias Adrian, Yakov Amihud, Hal Cole, Darrell Duffie, Bernard Dumas, Humberto Ennis, Mike Fleming, Nicolae Gârleanu, Ed Green, Joel Hasbrouck, Terry Hendershott, Jeremy Graveline, Narayana Kocherlakota, Anna Pavlova, Lasse Pedersen, Matt Richardson, Bill Silber, Stijn Van Nieuwerburgh, Neil Wallace, Robert Whitelaw, Randy Wright, seminar participants at the Federal Reserve Bank of Minneapolis, Federal Reserve Bank of Richmond, LSE, McGill, New Orleans, NY Fed, NYU, Oxford, UCLA Anderson, UCLA Economics, USC, Penn State, and participants at the American Finance Association 2005, Caesarea Center Annual Conference 2005, Federal Reserve Bank of Cleveland Summer Workshops in Money, Banking and Payments 2005, NBER Asset Pricing 2005, and Society for Economic Dynamics 2005 conferences for helpful comments. We are especially grateful to Mark Fisher, Kenneth Garbade, Tain Hsia-Schneider, and Frank Keane for discussions that greatly enhanced our understanding of the subject.

Submitted 30 October 2006

1 Introduction

In fixed-income markets some bonds trade at lower yields than others with almost identical cash flows. In the US, for example, just-issued ("on-the-run") Treasury bonds trade at lower yields than previously issued ("off-the-run") bonds maturing on nearby dates. Warga (1992) reports that an on-the-run portfolio returns on average 55bp below an off-the-run portfolio with matched duration. Similar phenomena exist in other countries. In Japan, for example, one "benchmark" government bond trades at a yield of 60bp below other bonds with comparable characteristics.[1]

How can the yields of bonds with almost identical cash flows differ by more than 50bp? Financial economists have suggested two apparently distinct hypotheses. First, on-the-run bonds are more valuable because they are significantly more liquid than their off-the-run counterparts. Second, on-the-run bonds constitute better collateral for borrowing money in the repo market. Namely, loans collateralized by on-the-run bonds offer lower interest rates than their off-the-run counterparts, a phenomenon referred to as "specialness."[2] These hypotheses, however, can provide only a partial explanation of the on-the-run phenomenon: one must still explain why assets with almost identical cash flows can differ in liquidity and specialness.

In this paper we propose a theory of the on-the-run phenomenon. We argue that liquidity and specialness are not independent explanations of this phenomenon, but can be explained simultaneously by short-selling activity. We determine liquidity and specialness endogenously, explain why they can differ across otherwise identical assets, and study their effect on prices. A calibration of our model for plausible parameter values can generate effects of the observed magnitude.

We consider an infinite-horizon steady-state economy with two assets paying identical cash flows. There is a continuum of agents experiencing transitory needs to hold long or short positions. An agent needing to be long buys an asset, and sells it later when the need disappears. Conversely, an agent needing to be short borrows an asset, sells it, and when the need disappears buys the asset back and delivers it to the lender. Trade involves two markets: a spot market to buy and sell, and a repo market where short-sellers can borrow assets. We assume that both markets operate through search, and model them as in the standard framework (e.g., Diamond (1982)) where agents are

[1] For US evidence, see also Amihud and Mendelson (1991), Krishnamurthy (2002), Goldreich, Hanke, and Nath (2002), and Strebulaev (2002). For Japan, see Mason (1987), Boudoukh and Whitelaw (1991), and Boudoukh and Whitelaw (1993).

[2] On liquidity, Sundaresan (2002) reports that trading volume of on-the-run bonds is about ten times larger than that of off-the-run bonds, and Fleming (2002) reports that bid-ask spreads of off-the-run bills are about five times larger than when these bills are on-the-run. Specialness is measured by comparing a bond's repo rate, which is the interest rate on a loan collateralized by the bond, to the general collateral rate, which is the highest quoted repo rate. Duffie (1996) reports an average specialness of 66bp for on-the-run bonds and 26bp for their off-the-run counterparts.

matched randomly over time in pairs and bargain over the terms of trade. This captures the over-the-counter structure of government-bond markets: transactions between dealers and customers are negotiated bilaterally over the phone, and dealers often negotiate bilaterally in the inter-dealer market.[3] Of course, the search framework is a stylized representation of government-bond markets - but so is the Walrasian auction which assumes multilateral trading. As long as search times are short, as is the case in our calibration, it is not obvious which model describes the markets better.

Our model has an asymmetric equilibrium in which assets trade at different prices despite the identical cash flows. The intuition is as follows. Suppose that all short-sellers prefer to borrow a specific asset. Because they initially sell and eventually buy the asset back, they increase the asset's trading volume in the spot market. This increases the asset's liquidity by reducing search frictions: with more volume, buyers and sellers become easier to locate. What makes short-sellers' concentration self-fulfilling is the constraint that they must deliver the same asset they borrowed. This constraint implies that a short-seller finds it optimal to borrow the asset that is easier to locate, which is precisely the asset that other short-sellers are borrowing.[4] The asset in which short-sellers concentrate trades at a premium for two reasons. Since it has a larger pool of buyers, it is easier to sell, and thus carries a liquidity premium. It also carries a specialness premium because its owners can lend it to short-sellers for a fee.

Our mechanism relies critically on short-sellers: we show that in their absence, assets trade at the same price. One could conjecture that even without short-sellers, asymmetric liquidity can arise in a self-fulfilling manner: one asset is harder to sell because its lack of liquidity drives buyers away. What rules out such asymmetries is that the difficulty to sell hurts sellers more than buyers because for buyers it becomes relevant only later in time when they turn into sellers. Thus, sellers of a less liquid asset are willing to lower the price enough to compensate buyers. But then buyers buy both assets, implying that both are equally easy to sell and trade at the same price.

Short-sellers can introduce asymmetries because, unlike longs, they are constrained to buy a specific asset - the one they borrowed. The mere presence of short-sellers, however, does not guarantee asymmetries because they could borrow both assets equally. Asymmetries are possible

[3] In the US, inter-dealer trading is conducted through brokers. Some brokers operate automated trading systems, structured as electronic limit-order books. Other brokers, however, operate voice-based systems in which orders are negotiated over the phone. Barclay, Hendershott, and Kotz (2006) report that automated systems account for about 85% of trading volume for on-the-run bonds, but the situation is reversed for off-the-run bonds. To explain this phenomenon, they propose a search-based model.

[4] The delivery constraint is standard in repo markets: lenders insist on receiving back the same asset they lent because of considerations of book-keeping, capital-gains taxes, etc. The prevalence of the delivery constraint is illustrated by the incidence of short-squeezes, whereby short-sellers have difficulty delivering the asset they borrowed and the asset's specialness in the repo market increases dramatically. For a description of short-squeezes see, for example, Dupont and Sack (1999).

because of the assumption of spot-market search. Indeed, because search generates a positive relationship between trading volume and liquidity, it implies that short-sellers have a preference for an asset that other short-sellers are borrowing. To emphasize the critical role of search, we show that if the spot market is Walrasian, then assets trade at the same price.

While the combination of short-sellers and spot-market search generates asymmetric liquidity, repo-market search ensures that the asymmetry can translate to a quantitatively significant price difference. Indeed, search precludes Bertrand competition between lenders in the repo market, and generates a positive lending fee. A positive fee gives rise to the specialness premium, which adds to the liquidity premium. Furthermore, the shorting costs implicit in the fee prevent arbitrageurs from eliminating the price difference between the two assets.

A calibration of our model can generate price effects of the observed magnitude even for very short search times. We show that the liquidity premium is small, and the effects are mostly generated by the specialness premium. Of course, this does not mean that liquidity does not matter; it rather means that liquidity can have large effects because it induces short-seller concentration and creates specialness.

Summarizing, our main contribution is to explain why assets with almost identical payoffs, such as on- and off-the-run bonds, can trade at significantly different prices. Our model also provides a framework for understanding other puzzling aspects of the on-the-run phenomenon. One apparent puzzle is that off-the-run bonds are viewed by traders as "scarce" and hard to locate, while at the same time being cheaper than on-the-run bonds. In our model, off-the-run bonds are indeed scarce from the viewpoint of short-sellers searching to buy and deliver them. Because, however, scarcity drives short-sellers away from these bonds, it makes them less liquid and less attractive to marginal buyers who are the agents seeking to establish long positions. Our theory also has the counter-intuitive implication that the trading activity of short-sellers can raise, rather than lower, an asset's price. This is because short-sellers increase both the asset's liquidity and specialness.

While our theory can explain price differences between on- and off-the-run bonds, it does not explain why short-sellers are more likely to concentrate in on-the-run bonds.[5] We show, however, that if assets differ enough in their supplies (i.e., issue sizes), the equilibrium becomes unique with short-sellers concentrating in the largest-supply asset. This is consistent with the commonly held view that off-the-run bonds are in smaller effective supply (because, e.g., they become "locked away" in the portfolios of buy-and-hold investors). Of course, our theory cannot address the decrease in

[5]Our multiple-equilibria view seems to fit the Japanese government-bond market: Boudoukh and Whitelaw (1991) and Boudoukh and Whitelaw (1993) show that the benchmark bond in which liquidity concentrates can be arbitrary.

effective supply because it assumes a steady state.

This paper is closely related to Duffie's (1996) theory of repo specialness. In Duffie, short-sellers need to borrow an asset and sell it in a market with exogenous transaction costs. Assets differ in transaction costs, and those with low costs are on special because they are in high demand by short-sellers. The main difference with Duffie is that instead of explaining specialness taking liquidity (transaction costs) as exogenous, we explain why both liquidity and specialness can differ for otherwise identical assets. Krishnamurthy (2002) proposes a model building on Duffie (1996) that links the specialness premium to an exogenous liquidity premium. This link is also present in our model where the liquidity premium is endogenous.[6]

Duffie, Gârleanu and Pedersen introduce search and matching in models of dynamic asset market equilibrium.[7] In Duffie, Gârleanu, and Pedersen (2006) investors seek to establish long positions, and in Duffie, Gârleanu, and Pedersen (2005) trade is intermediated through dealers. Duffie, Gârleanu, and Pedersen (2002) model search in the repo market and show that it generates a positive lending fee. Our focus differs in that we seek to explain price differences among otherwise identical assets. This leads us to consider a multi-asset model while they assume only one asset, and allow for search in both the spot and the repo market.

Vayanos and Wang (2006) and Weill (2004) develop multi-asset models with search, in which assets with identical payoffs can trade at different prices. They assume no short-sellers, however, and the price differences are driven by the constraint that longs must choose which asset to buy before starting the search process. This constraint is somewhat implausible in the context of the Treasury market since, for example, longs have the flexibility of buying any asset when they contact a dealer. In the present paper, by contrast, price differences are driven by the more standard constraint that short-sellers must deliver the same asset they borrowed. Furthermore, the presence of short-sellers allows us to explore the interplay between liquidity and specialness, and generate much larger price effects.

This paper is related to the monetary-search literature building on Kiyotaki and Wright (1989) and Trejos and Wright (1995). Aiyagari, Wallace, and Wright (1996) provide an example of an economy in which fiat monies (intrinsically worthless and unbacked pieces of paper) endogenously differ in their price and liquidity. Wallace (2000) analyzes the relative liquidity of currency and

[6]Empirical studies by Cornell and Shapiro (1989), Jordan and Jordan (1997), Buraschi and Menini (2002), Krishnamurthy (2002), Graveline and McBrady (2004), and Moulton (2004) show that on-the-run bond prices contain specialness premia consistent with Duffie (1996) and our model. We return to these studies in Section 4.6.

[7]See also Burdett and O'Hara (1987) and Keim and Madhavan (1996) for search-theoretic models of block trading in the upstairs market.

dividend-paying assets in a model based on asset indivisibility. Our relative contribution is to compare dividend-paying assets as opposed to currency, and introduce short sales.

This paper is also related to the literature on equilibrium asset pricing with transaction costs. (See, for example, Amihud and Mendelson (1986), Constantinides (1986), Aiyagari and Gertler (1991), Heaton and Lucas (1996), Vayanos (1998), Vayanos and Vila (1999), Huang (2003), and Lo, Mamaysky, and Wang (2004).) We add to that literature by endogenizing transaction costs.

Pagano (1989) generates asymmetric liquidity because traders can concentrate in one of multiple markets.[8] Our work differs because we consider concentration across assets rather than market venues for the same asset. Boudoukh and Whitelaw (1993) show that asymmetric liquidity can arise when a monopolistic bond issuer uses liquidity as a price-discrimination tool.

The rest of this paper is organized as follows. Section 2 presents the model. Section 3 shows that the model is based on a minimum set of assumptions: when any assumption is relaxed, the Law of One Price holds. Section 4 contains our main model and results. Section 5 calibrates the model and Section 6 concludes. All proofs are in the Appendix.

2 Model

Time is continuous and goes from zero to infinity. There are two assets $i \in \{1, 2\}$ that pay an identical dividend flow δ and are in identical supply S. Agents are infinitely lived and form a continuum with infinite mass. They can hold long or short positions in either asset. For simplicity, however, we allow for only three types of portfolios: long one share (of either asset), short one share, or no position.

Agents derive a utility flow from holding a position. The utility flow is zero for an agent holding no position. An agent holding $q \in \{-1, 1\}$ shares of either asset derives utility flow $q(\delta + x_t) - y$, where $y > 0$ and $\{x_t\}_{t \geq 0}$ is a stochastic process taking the values $\overline{x} > 0$, 0, and $-\underline{x} < 0$. We refer to agents with $x_t = \overline{x}$ as high-valuation, $x_t = 0$ as average-valuation, and $x_t = -\underline{x}$ as low-valuation. Agents' lifetime utility is the present value (PV) of expected utility flows, net of payments for asset transactions, and discounted at a rate $r > 0$.

Our utility specification can be interpreted in terms of risk aversion. If the parameter δ is an

[8] See also Ellison and Fudenberg (2003) for a general analysis of the coexistence of markets, and Economides and Siow (1988) for a spatial model of market formation. See also Admati and Pfleiderer (1988) and Chowdhry and Nanda (1991) for models where trading is concentrated in a specific time or location because of asymmetric information.

expected rather than actual dividend flow, $q\delta$ represents a position's expected cash flow. This cash flow needs to be adjusted for risk. The parameter y represents a cost of risk bearing, which is positive for both long and short positions. The parameters $\overline{x}$ and $\underline{x}$ represent hedging benefits. For example, low-valuation agents could be hedging the risk of a long position held in a different but correlated market. A short position would give these hedgers an extra utility $\underline{x}$, while a long position would give them a disutility $-\underline{x}$.[9] In Appendix E we derive our utility specification from first principles.[10] We assume that agents have CARA preferences over a single consumption good, and can invest in a riskless asset with return r and in two identical risky assets with expected dividend flow δ. Moreover, agents receive a random endowment whose correlation with the dividend flow can be positive (low-valuation), zero (average-valuation), or negative (high-valuation). These assumptions give rise to our reduced-form specification, with the parameters y, $\overline{x}$, $\underline{x}$ being functions of the agents' risk-aversion, the variance of the dividend flow, and the endowment correlation. We leave the CARA specification to the Appendix because the reduced form conveys the main intuitions without burdening the derivations.

At each point in time, there is a flow $\overline{F}$ of average-valuation agents who switch to high valuation, and a flow $\underline{F}$ who switch to low valuation. Conversely, high-valuation agents revert to average valuation with Poisson intensity $\overline{\kappa}$, and low-valuation agents do the same with Poisson intensity $\underline{\kappa}$. Thus, the steady-state measures of high- and low-valuation agents are $\overline{F}/\overline{\kappa}$ and $\underline{F}/\underline{\kappa}$, respectively. Given that the measure of average-valuation agents is infinite, an individual agent's switching intensity from average to high or low valuation is zero.

For simplicity, we impose the following parameter restrictions.

Assumption 1. $\overline{x} + \underline{x} > 2y > \overline{x}$.

Assumption 2. $\frac{\overline{F}}{\overline{\kappa}} > 2S + \frac{\underline{F}}{\underline{\kappa}}$.

Assumption 1 ensures that low-valuation agents are willing to short-sell in equilibrium, while average-valuation agents are not. Indeed, consider a low-valuation agent who establishes a short position with a high-valuation agent as the long counterpart. The flow surplus of the transaction is the sum of the high-valuation agent's utility flow from the long position plus the low-valuation agent's utility flow from the short position:

$$[\delta + \overline{x} - y] + [-(\delta - \underline{x}) - y] = \overline{x} + \underline{x} - 2y.$$

[9] Hedgers in the Treasury market can be, for example, dealers in corporate bonds or mortgage-backed securities, who need to hedge the interest-rate risk of their inventories. For a discussion of hedging in the Treasury market, see Dupont and Sack (1999).

[10] See Duffie, Gârleanu, and Pedersen (2006) for a similar derivation.

Assumption 1 ensures that this is positive because the combined hedging benefits $\overline{x} + \underline{x}$ exceed the total cost $2y$ of risk bearing. On the other hand, the flow surplus when the short-seller is an average-valuation agent is $[\delta + \overline{x} - y] + [-\delta - y] = \overline{x} - 2y < 0$.

Assumption 2 ensures that high-valuation agents are the marginal asset holders. Indeed, the aggregate asset supply is the sum of the supply $2S$ from the issuers plus the supply from the short-sellers. Since low-valuation agents are the only short-sellers and short one share, the latter supply is equal to their measure $\underline{F}/\underline{\kappa}$. The aggregate supply is thus smaller than the measure $\overline{F}/\overline{\kappa}$ of high-valuation agents, meaning that these agents are marginal.

In what follows, we focus on steady-state equilibria. Assumptions 1 and 2 ensure that in such equilibria high-valuation agents seek to establish long positions, low-valuation agents seek to establish short positions, and average-valuation agents stay out of the market.

3 Market Settings Consistent with the Law of One Price

In our main model of Section 4 there are two markets, both operating through search: a spot market to buy and sell assets, and a repo market where short-sellers can borrow assets. In this section we take a step back and argue that the combination of short-sellers and a search spot market are necessary for explaining the on-the-run phenomenon. Namely, we consider benchmark settings where either short-sales are not allowed or the spot market is Walrasian. We show that in these settings the Law of One Price holds, i.e., assets 1 and 2 trade at the same price.

3.1 No Short-Sales

We start with the case where short-sales are not allowed. The repo market is then shut, and agents trade only in the spot market. Not surprisingly, the Law of One Price holds when the spot market is Walrasian.

Proposition 1 (No Short-Sales, Walrasian Spot Market). *Suppose that short-sales are not allowed. In a Walrasian equilibrium both assets trade at the same price*

$$p = \frac{\delta + \overline{x} - y}{r}.$$

Moreover, high-valuation agents buy one share or stay out of the market, and low- and average-valuation agents stay out of the market.

The intuition why both assets trade at the same price is straightforward: if one were cheaper, it would be the only one demanded by agents. The common price of the assets is determined by the marginal holders. From Assumption 2, these are the high-valuation agents, and the price is equal to the PV of their utility flow $\delta + \overline{x} - y$ from holding one share. Under this price, high-valuation agents are indifferent between buying and staying out of the market, while other agents prefer to stay out of the market.

We next assume that the spot market operates through search. As in the standard search framework, we assume that buyers and sellers are matched randomly over time in pairs. The buyers are high-valuation agents, and the sellers are average-valuation agents who bought when they were high-valuation. We denote by $\mu_{\overline{b}}$ the measure of buyers and by $\mu_{\overline{s}i}$ the measure of sellers of asset i.

We assume that an agent establishes contact with others at Poisson arrival times with fixed intensity. Moreover, there is random matching in that conditional on establishing a contact, all agents are "equally likely" to be contacted. Thus, an agent meets members of a given group with Poisson intensity proportional to that group's measure. For example, a buyer meets sellers of asset i with Poisson intensity $\lambda\mu_{\overline{s}i}$, where λ is a parameter measuring the efficiency of search. The Law of Large Numbers (see Duffie and Sun (2004)) implies that meetings between buyers and sellers of asset i occur at a deterministic rate $\lambda\mu_{\overline{b}}\mu_{\overline{s}i}$.

When a buyer meets a seller, they bargain over the price. We assume that bargaining is efficient, in that trade occurs whenever the buyer's reservation utility exceeds the seller's. If trade occurs, the price is set so that the buyer receives a fraction $\phi \in [0, 1]$ of the surplus.

Proposition 2 (No Short-Sales, Search Spot Market). *Suppose that short-sales are not allowed. In a search equilibrium all buyer-seller meetings result in a trade, and both assets trade at the same price.*

Proposition 2 shows that the Law of One Price holds even in the presence of search frictions. In particular, there do not exist asymmetric equilibria in which assets differ in liquidity. One could conjecture, for example, an equilibrium in which buyers refuse to trade when they meet sellers of asset 2, preferring to wait for sellers of asset 1. This behavior could be based on a self-fulfilling expectation of low liquidity: a buyer fears that asset 2 will be difficult to sell because he expects that other buyers will also refuse to buy. What rules out such equilibria is that the difficulty to sell hurts sellers even more than buyers because for buyers it becomes relevant only later in time when they turn into sellers. As a consequence, sellers of asset 2 are willing to lower the price enough to

compensate buyers for any difficulties they will encounter when selling the asset. But then both assets have the same buyer pool, consisting of high-valuation agents. Therefore, they are equally easy to sell, and trade at the same price.

Proposition 2 implies that search frictions alone are not enough to generate price differences among otherwise identical assets. One must also explain why assets' buyer pools can be different. Vayanos and Wang (2006) and Weill (2004) derive price differences in settings where buyers must choose which asset to buy before starting the search process.[11] This constraint, however, is somewhat implausible in the context of the Treasury market. Suppose, for example, that a buyer contacts a dealer for an on-the-run bond. If the dealer happens to have an attractively priced off-the-run bond in inventory, nothing prevents the buyer from switching to that bond. The constraint becomes much more plausible if buyers are not agents seeking to initiate long positions (as in Vayanos and Wang (2006) and Weill (2004)), but seeking to cover previously established short positions. Indeed, these agents must deliver the same asset they borrowed. Considering short-sellers and the related issue of repo specialness is a central and novel element in our theory.

3.2 Short-Sales – Walrasian Spot Market

We next allow for short-sales but assume that the spot market is Walrasian. To motivate our modelling of the repo market, we recall the mechanics of repo transactions. In a repo transaction a lender turns his asset to a borrower in exchange for cash. At maturity the borrower must return an asset from the same issue, and the lender returns the cash together with some previously-agreed interest-rate payment, called the repo rate. Hence, a repo transaction is effectively a loan of cash collateralized by the asset. Treasury securities differ in their repo rates. Most of them share the same rate, called the general collateral rate, which is the highest quoted repo rate and is close to the Fed Funds Rate. The specialness of an asset is defined as the difference between the general collateral rate and its repo rate. In our model, instead of assuming that the lender pays a low repo rate to the borrower, we assume that the borrower pays a positive flow fee w to the lender. Hence, the implied repo rate is the difference $r - w/p$ between the risk-free rate and the lending fee per dollar, and the specialness is simply w/p.

When the spot and the repo market are both Walrasian, the Law of One Price holds in both

[11] The price differences in these papers compensate buyers for the difficulty in locating an asset. Suppose, for example, that asset 2 is harder to locate than asset 1. Then, no buyer will search for asset 2 if its price is greater or equal than asset 1's. If, however, searching for asset 2 does not preclude a simultaneous search for asset 1, price differences disappear (as shown in Proposition 2).

markets: the assets trade at the same price and carry the same lending fee. Furthermore, the fee is zero. Indeed, with a positive fee, agents would prefer to lend their assets in the repo market rather than holding them. This would be inconsistent with equilibrium since assets are in positive supply.

Proposition 3 (Short-Sales, Walrasian Spot and Repo Markets). *Suppose that short-sales are allowed. In a Walrasian equilibrium both assets trade at the same price*

$$p = \frac{\delta + \overline{x} - y}{r}$$

and the lending fee w is zero. Moreover, high-valuation agents buy one share or stay out of the market, low-valuation agents short one share, and average-valuation agents stay out of the market.

We next introduce search in the repo market, while keeping the spot market Walrasian. We assume that lenders and borrowers are matched randomly over time in pairs. The lenders are high-valuation agents owning an asset, and the borrowers are low-valuation agents seeking to initiate a short-sale. After borrowing an asset, a low-valuation agent sells it immediately in the Walrasian spot market. When the agent reverts to average valuation, she buys the asset back and returns it to the lender, who then searches for a new borrower. If the lender reverts to average valuation while the asset is on loan, he can choose to terminate the repo contract. In that case, the borrower buys the asset in the spot market and returns it to the lender. The lender then sells the asset, while the borrower searches for a new lender. We denote by $\mu_{\underline{bo}}$ the measure of borrowers and by $\mu_{\overline{\ell}i}$ the measure of lenders of asset i. We assume the same matching technology as in Section 3.1: meetings between borrowers and lenders of asset i occur at the deterministic rate $\nu\mu_{\underline{bo}}\mu_{\overline{\ell}i}$, where ν is a parameter measuring the efficiency of repo-market search.

When a borrower meets a lender, they bargain over the lending fee. We assume that bargaining is efficient, in that the repo transaction occurs whenever there is a positive surplus. If the transaction occurs, the lending fee is set so that the lender receives a fraction $\theta \in [0, 1]$ of the surplus.

Proposition 4 (Short-Sales, Walrasian Spot Market, Search Repo market). *Suppose that short-sales are allowed, the spot market is Walrasian, and the repo market operates through search. In equilibrium both assets trade at the same price and carry the same positive lending fee.*

Proposition 4 implies that search frictions in the repo market alone cannot generate departures from the Law of One Price: the assets trade at the same price and carry the same lending fee. The only effect of repo-market frictions is that the fee is positive. The mechanism is the same as

in Duffie, Gârleanu, and Pedersen (2002): search precludes Bertrand competition between lenders because borrowers can only meet one lender at a time.

To explain why the Law of One Price holds, consider a possible asymmetric equilibrium where short-sellers refuse to borrow asset 2, preferring to wait for a lender of asset 1. Such behavior could be based on the expectation that asset 2 might be harder to deliver when unwinding the repo contract. But with a Walrasian spot market, both assets can be costlessly bought and delivered. Therefore, short-sellers are willing to borrow both. Note that the same conclusion would hold if there are transaction costs in the spot market, provided that these are equal across assets.

While search frictions in the repo market alone cannot explain the on-the-run puzzle, they can be part of the explanation. Indeed, suppose that for some (yet unexplained) reason, short-sellers prefer to borrow a specific asset, e.g., asset 1. Then, the lenders of asset 1 can negotiate a positive lending fee, while there is no fee for asset 2. Since the lending fee constitutes an additional cash flow derived from an asset, it raises the price of asset 1 above that of asset 2, resulting in a departure from the Law of One Price.

Why might short-sellers prefer to borrow a specific asset? A natural reason is that the asset is easier to deliver because of lower transaction costs in the spot market. This is very plausible in the context of Treasuries: locating a large quantity of a specific off-the-run issue can be harder than for on-the-run issues. One must explain, however, why transaction costs can differ across two otherwise identical assets. As we argue in the next section, a natural explanation, and one which is central to our theory, is based on search frictions in the spot market.[12]

4 Departing from the Law of One Price

Our theory of the on-the-run phenomenon is based on short-sellers and search frictions in the spot and the repo market. In a nutshell, the mechanism is as follows. Suppose that all short-sellers prefer to borrow a specific asset. Because they initially sell and eventually buy the asset back, they increase the asset's trading volume in the spot market. This increases the asset's liquidity by reducing search frictions: with more volume, buyers and sellers become easier to locate. The

[12] Our analysis assumes that the only delivery method is to buy the asset in the spot market. An alternative method is through the repo market: short-sellers could borrow the asset from a new lender and deliver it to the original one. (This method is relevant only when short-sellers wish to maintain the short position and it is the lender who needs to sell.) One could argue that the ease of delivering an asset has to do with the repo market, and not with transaction costs in the spot market. In our model, however, both assets are equally easy to locate in the repo market because they are in equal supply and thus have the same measure of lenders. Moreover, in practice, while it might be easier to locate an on- rather than an off-the-run bond in the repo market, such differences are perceived to be of secondary importance relative to the corresponding differences in the spot market.

increase in liquidity is, in turn, what makes the asset attractive to short-sellers because they can unwind their positions more easily. The asset in which short-sellers concentrate trades at a premium for two reasons. Since it has a larger pool of buyers, it is easier to sell, and thus carries a liquidity premium. It also carries a specialness premium because its owners can lend it to short-sellers for a fee.

The interaction between short-sellers and spot-market search is at the heart of our theory. Search can generate differences in spot-market liquidity among otherwise identical assets, but only if some investors trade one asset more than the other. As shown in Proposition 2, such asymmetric trading is hard to rationalize with longs: since they have the flexibility to buy either asset, they constitute a common buyer pool for both assets, and trade them equally. Short-sellers, by contrast, are constrained to buy the same asset they borrowed, and thus can generate asymmetric trading if they have a preference for a specific asset. This preference can arise if one asset is easier to deliver than the other. As shown in Proposition 4, such differences across assets are hard to rationalize without differences in spot-market liquidity, which is precisely what search can generate.

While the combination of short-sellers and spot-market search generates asymmetric liquidity, repo-market search ensures that the asymmetry can translate to a quantitatively significant price difference. Indeed, with a Walrasian repo market, both assets would have a lending fee of zero. Therefore, there would be no specialness premium–which according to our calibration is significantly larger than the liquidity premium. Moreover, a zero lending fee would imply no shorting costs. Thus, arbitrageurs could profit from (and eventually eliminate) the liquidity premium by selling the more liquid asset and buying the less liquid one. In most of our analysis we do not consider arbitrage strategies because we restrict agents to hold either long or short positions. In Section 4.4, however, we allow for such strategies and show that they can be unprofitable in the presence of repo-market frictions.

In the rest of this section, we develop our theory formally and show our main results. Because the theory involves search in both the spot and the repo market, the analysis is more complicated than in Section 3. In particular, we need to describe carefully the various types of agents (e.g., buyers, sellers, lenders, borrowers), the transitions between types, the bargaining outcome when two agents meet, and the concept of equilibrium in the search markets.

4.1 Agent Types, Transitions, and Population Measures

In this section we describe the types of agents and the transitions between types. The transitions are partly generated by agents' trading strategies, whose optimality we defer to Section 4.3. Figure 1 summarizes agents' life-cycles. The top part of the figure concerns a high-valuation agent who is initially a buyer $\overline{b}$, seeking a seller of either asset in the spot market. If he reverts to average valuation before meeting a seller, he exits the market. Otherwise, if he meets a seller of asset $i \in \{1, 2\}$, he bargains over the price p_i and buys the asset. He then becomes a lender $\overline{\ell} i$ of asset i in the repo market, seeking a borrower. If he reverts to average valuation before meeting a borrower, he exits the repo market and becomes a seller $\overline{s} i$ of asset i in the spot market. Upon meeting a buyer, he bargains over the price p_i, sells the asset, and exits the market. If instead the lender $\overline{\ell} i$ meets a borrower and there are gains from trade, he bargains over the lending fee w_i and enters in a repo contract. We describe the states within a repo contract shortly.

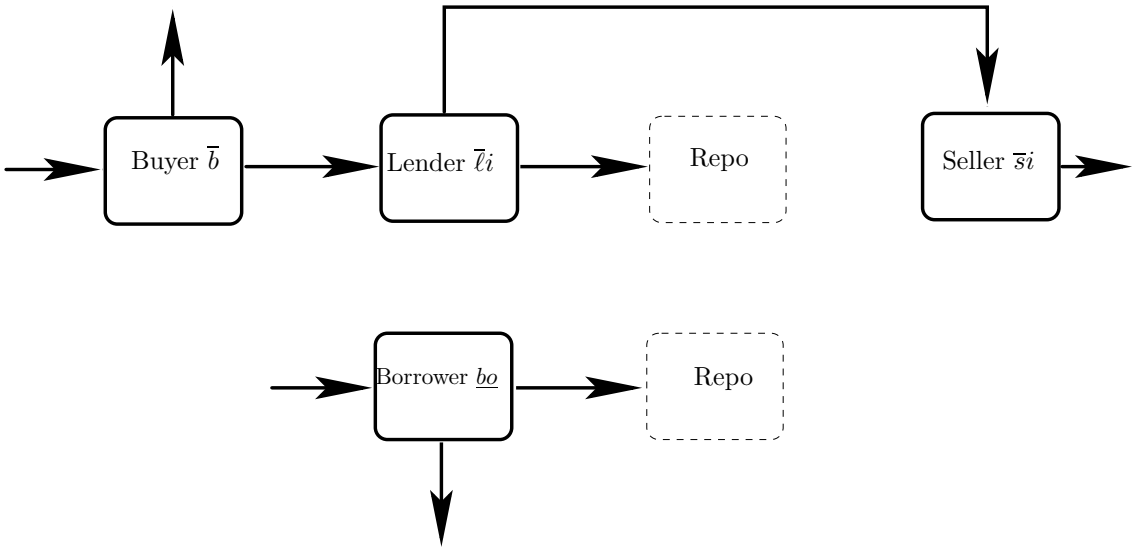

Figure 1: Life-cycles. Figure 2 magnifies the two repo boxes.

The bottom part of Figure 1 concerns a low-valuation agent who is initially a borrower $\underline{bo}$, seeking a lender in the repo market. If she reverts to average valuation before meeting a lender, she exits the market. Otherwise, if she meets a lender of asset i and there are gains from trade, she bargains over the lending fee w_i and enters in a repo contract.

Figure 2 describes the states within a repo contract. The agent borrowing the asset, represented by the lower dashed box, can be of three types: $\underline{s} i$, $\underline{n} i$, and $\underline{b} i$. Initially, she is a seller $\underline{s} i$, seeking to sell asset i in the spot market. After selling the asset, she becomes a non-searcher $\underline{n} i$. Finally,

when she reverts to average valuation, she seeks to unwind her short position and becomes a buyer $\underline{b}i$ of asset i in the spot market. The agent lending the asset, represented by the upper dashed box, can be of three types, depending on the current type of his borrower. For example, type $\overline{n}\underline{s}i$ is a non-searcher whose repo counterparty is in state $\underline{s}i$.

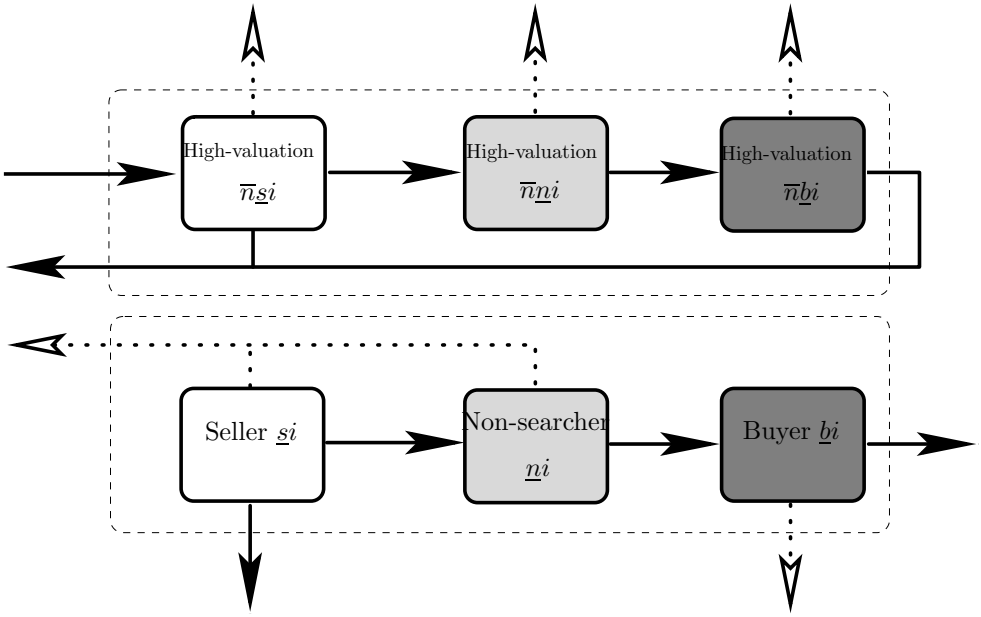

Figure 2: The states of a lender and a borrower within a repo contract.

A repo contract can be terminated by either the borrower or the lender, but in different ways. The borrower can terminate by delivering the same asset she borrowed, while the lender can terminate by asking for instant delivery. Terminations are described by the arrows leaving the dashed boxes, with solid arrows corresponding to borrower-driven terminations, and dotted arrows to lender-driven ones.

A borrower terminates the contract when she is a buyer $\underline{b}i$ and meets a seller. She can also terminate when she reverts to average valuation before selling the asset, i.e., while being a seller $\underline{s}i$. In both cases, she delivers the asset and exits the market, while the lender returns to the pool $\overline{\ell}i$ of lenders.

A lender terminates the contract when he reverts to average valuation. If the borrower has the asset in hand because she is of type $\underline{s}i$, she delivers it instantly. The lender then becomes a seller $\overline{s}i$, while the borrower returns to the pool $\underline{b}o$ of borrowers. If the borrower does not have the asset because she sold it and is of type $\underline{n}i$ or $\underline{b}i$, instant delivery is impossible because of search. In that event, we assume that the lender seizes some cash collateral previously posted by the borrower and

exits the market.[13] The borrower returns to the pool $\underline{bo}$ of borrowers if she still wishes to hold a short position (i.e., is of type $\underline{n}i$), and exits the market otherwise.

We refer to the different states in agents' life-cycles as "types." The possible types are $\overline{b}$ and $\{\overline{\ell}i, \overline{n}\underline{s}i, \overline{n}\underline{n}i, \overline{n}\underline{b}i, \overline{s}i\}_{i\in\{1,2\}}$ for high-valuation agents, and $\underline{bo}$ and $\{\underline{s}i, \underline{n}i, \underline{b}i\}_{i\in\{1,2\}}$ for low-valuation agents. We denote by $\mathcal{T}$ the set of types, and by μ_τ the measure of agents of type $\tau \in \mathcal{T}$. Finally, we denote by bi the group of all buyers of asset i (both high- and low-valuation), and by si the group of all sellers. The measures μ_{bi} and μ_{si} of these groups are

$$\mu_{bi} = \mu_{\overline{b}} + \mu_{\underline{b}i} \tag{1}$$

$$\mu_{si} = \mu_{\overline{s}i} + \mu_{\underline{s}i}. \tag{2}$$

The steady-state measures must satisfy the market-clearing equation

$$\mu_{\overline{\ell}i} + \mu_{si} = S \tag{3}$$

since assets are held by either lenders or sellers, and

$$\mu_{\overline{n}i} \equiv \mu_{\overline{n}\underline{s}i} + \mu_{\overline{n}\underline{n}i} + \mu_{\overline{n}\underline{b}i} = \mu_{\underline{s}i} + \mu_{\underline{n}i} + \mu_{\underline{b}i} \tag{4}$$

since there must be an equal measure of high- and low-valuation agents involved in repo contracts. Additionally, the steady-state measures must be such that the inflow into a type is equal to the outflow. For example, the inflow into the type $\overline{b}$ of high-valuation buyers is $\overline{F}$ because of the new entrants. The outflow is the sum of $\overline{\kappa}\mu_{\overline{b}}$ because some buyers revert to average valuation and exit the market, and $\sum_{i=1}^{2} \lambda\mu_{si}\mu_{\overline{b}}$ because some buyers meet with sellers. (We are assuming the same matching technology as in Section 3.) Therefore, the inflow-outflow equation for type $\overline{b}$ is

$$\overline{F} = \overline{\kappa}\mu_{\overline{b}} + \sum_{i=1}^{2} \lambda\mu_{si}\mu_{\overline{b}}. \tag{5}$$

In Appendix B we derive the remaining inflow-outflow equations, and show that the resulting system determines uniquely the steady-state measures of all types.

[13] This assumption is for simplicity. An alternative assumption is that the borrower can search for the asset under a late-delivery penalty, but this would not change the basic intuitions.

In Appendix C we show that because collateral acts as a transfer, its specific value does not affect any equilibrium variable except the price of the repo contract: high-valuation agents accept to lend their asset for a lower fee if they can seize more collateral. To downplay this effect, we set the collateral equal to the utility of a seller $\overline{s}i$. This ensures that upon reverting to average valuation, agent $\overline{n}i$ is equally well off when receiving the asset (thus becoming a seller $\overline{s}i$) or the cash collateral.

4.2 Bargaining

Prices are the outcome of pairwise bargaining between buyers and sellers, and lending fees the outcome of bargaining between borrowers and lenders. To determine these outcomes, we need to compute the utility V_τ of being type $\tau \in \mathcal{T}$. As is standard in the search literature, the flow value rV_τ of being type τ can be derived from the flow benefits accruing to that type plus the transitions to other types. (The transitions are partly generated by agents' trading strategies, whose optimality we examine in Section 4.3.) The flow value of being a high-valuation buyer $\overline{b}$, for example, is

$$rV_{\overline{b}} = -\overline{\kappa}V_{\overline{b}} + \sum_{i=1}^{2} \lambda\mu_{si}(V_{\overline{\ell}i} - p_i - V_{\overline{b}}), \tag{6}$$

because the flow benefits are zero and the transitions are (i) revert to average valuation at rate $\overline{\kappa}$ and exit the market (utility zero and net utility $-V_{\overline{b}}$), and (ii) meet a seller of asset $i \in \{1, 2\}$ at rate $\lambda\mu_{si}$, buy at price p_i, and become a lender $\overline{\ell}i$ (utility $V_{\overline{\ell}i}$ and net utility $V_{\overline{\ell}i} - p_i - V_{\overline{b}}$). We derive the equations for all other types in Appendix C.

We assume that bargaining in the repo market is as in Section 3.2: a repo transaction occurs whenever there is a positive surplus, and the lender receives a fraction $\theta \in [0, 1]$ of the surplus. Since a repo transaction turns the lender $\overline{\ell}i$ into type $\overline{n}\underline{s}i$, the lender's surplus is $V_{\overline{n}\underline{s}i} - V_{\overline{\ell}i}$. The total surplus is $\Sigma_i \equiv V_{\overline{n}\underline{s}i} - V_{\overline{\ell}i} + V_{\underline{s}i} - V_{\underline{bo}}$ because the borrower $\underline{bo}$ becomes a seller $\underline{s}i$. Therefore, the lending fee is implicitly determined by

$$V_{\overline{n}\underline{s}i} - V_{\overline{\ell}i} = \theta\Sigma_i = \theta(V_{\overline{n}\underline{s}i} - V_{\overline{\ell}i} + V_{\underline{s}i} - V_{\underline{bo}}). \tag{7}$$

Bargaining in the spot market is more complicated than in Section 3.1 because for each asset i there are two buyer types, $\overline{b}$ and $\underline{b}i$, and two seller types, $\overline{s}i$ and $\underline{s}i$. To ensure that the bargaining outcome is consistent with types being private information, we consider an explicit bargaining game (rather than assuming that each buyer type receives the same fraction ϕ of the surplus).[14] We assume that the buyer and the seller make simultaneous offers. If the offers generate a set of mutually acceptable prices, trade occurs at the mid-point of that set. Otherwise, the meeting ends and agents return to the search pool. Types' reservation values are as follows. Type $\overline{b}$ has reservation value $\Delta_{\overline{b}} \equiv V_{\overline{\ell}i} - V_{\overline{b}}$ because after buying the asset he becomes a lender with utility

[14] Private information is a natural assumption in our setting since agents' types depend on their portfolio positions and hedging needs, which are not observable to their counterparties.

$V_{\overline{\ell}i}$. Type $\underline{b}i$ has reservation value $\Delta_{\underline{b}i} \equiv -V_{\underline{b}i}$ because after buying she delivers the asset and exits the market. Likewise, the reservation values of the seller types are $\Delta_{\overline{s}i} \equiv V_{\overline{s}i}$ and $\Delta_{\underline{s}i} \equiv V_{\underline{s}i} - V_{\underline{n}i}$. Because type $\overline{b}$ receives a hedging benefit from holding the asset while type $\overline{s}i$ does not, reservation values satisfy $\Delta_{\overline{b}} > \Delta_{\overline{s}i}$. They also satisfy $\Delta_{\underline{b}i} > \Delta_{\underline{s}i}$ because type $\underline{s}i$ receives a hedging benefit from holding a short position while type $\underline{b}i$ does not. To complete the ranking, we assume for simplicity that short-sellers are the infra-marginal traders, both as sellers and as buyers, i.e.,[15]

$$\Delta_{\underline{b}i} > \Delta_{\overline{b}} > \Delta_{\overline{s}i} > \Delta_{\underline{s}i}. \tag{8}$$

We focus on simple equilibria of the bargaining game in which all types make the same offer p_i. This offer must be in $[\Delta_{\overline{s}i}, \Delta_{\overline{b}}]$ to ensure that all types realize a non-negative surplus. Given the buyer's strategy, asking p_i is optimal for a seller - a higher ask would preclude trading while a lower ask would lower the transaction price. Likewise, given the seller's strategy, bidding p_i is optimal for a buyer. Obviously any $p_i \in [\Delta_{\overline{s}i}, \Delta_{\overline{b}}]$ is an equilibrium. We do not select among these, but instead treat the parameter ϕ defined by

$$p_i = \phi\Delta_{\overline{s}i} + (1-\phi)\Delta_{\overline{b}} = \phi V_{\overline{s}i} + (1-\phi)(V_{\overline{\ell}i} - V_{\overline{b}}), \tag{9}$$

as exogenous. This parameter measures the buyers' bargaining power because it is equal to the fraction of the total surplus $\Delta_{\overline{b}} - \Delta_{\overline{s}i}$ that the marginal buyer $\overline{b}$ can extract.

4.3 Equilibrium

An equilibrium is characterized by (i) measures μ_τ for all agent types $\tau \in \mathcal{T}$, (ii) utilities V_τ for all agent types $\tau \in \mathcal{T}$, (iii) prices and lending fees (p_i, w_i) for $i \in \{1,2\}$, and (iv) short-selling decisions ν_i for $i \in \{1,2\}$, where $\nu_i \equiv \nu$ if low-valuation agents borrow asset i and $\nu_i \equiv 0$ otherwise. These variables solve a fixed-point problem. The measures are determined from Equations (3)-(5) and (B.1)-(B.6), as a function of the short-selling decisions. The utilities, prices, and lending fees are determined from Equations (6), (7), (9) and (C.1)-(C.9), as a function of the measures and short-selling decisions. Finally, the short-selling decisions are determined as a function of the utilities from

$$\nu_i = \nu \Leftrightarrow \Sigma_i \geq 0, \tag{10}$$

[15] This assumption makes the analysis more transparent because it ensures that marginal traders are comparable across assets even in equilibria where short-selling is concentrated in one asset. In Section 4.3 we show that Equation (8) is satisfied under appropriate restrictions on exogenous parameters.

i.e., agents short asset i if the surplus Σ_i associated to a repo transaction is positive.

A solution to the fixed-point problem is an equilibrium if it satisfies two additional requirements. First, the conjectured trading strategies must be optimal, i.e., high- and low-valuation agents must follow the strategies described in Section 4.1, and average-valuation agents must hold no position. Second, the buyers' and sellers' reservation values must be ordered as in Equation (8).

Computing an equilibrium can, in general, be done only numerically. Fortunately, however, closed-form solutions can be derived when search frictions are small, i.e., λ and ν are large.[16] In the remainder of this section we focus on this case, emphasizing the intuitions gained by the closed-form solutions. We complement our asymptotic analysis with a numerical calibration in Section 5. When search frictions are small, the measure of agents in the "short" side of a market goes to zero. The short side in the repo market are the borrowers because they enter the market at a flow rate, while the lenders are the asset-holders and constitute a stock. The short side in the spot market are the sellers because Assumption 2 ensures that the asset demand generated by high-valuation agents exceeds the asset supply generated by issuers and short-sellers.

Given that assets are symmetric, a natural equilibrium is one in which low-valuation agents borrow both assets. Propositions 5 and 6 show existence of such a symmetric equilibrium and determine its properties.

Proposition 5. *Suppose that*

$$\frac{\underline{x} + \frac{\kappa}{r+\overline{\kappa}+g_s}\overline{x}}{1 + \frac{\kappa}{r+\overline{\kappa}+g_s}} > 2y \tag{11}$$

and $\phi, \theta \neq 1$, where g_s is defined by Equation (B.28) of Appendix B. Then, for large λ and ν, there exists a symmetric equilibrium in which low-valuation agents borrow both assets. Prices, lending fees, and population measures are identical across assets.

For simplicity, we suppress the asset subscript for variables in the symmetric equilibrium. In the proof of Proposition 5 we confirm that the measures of sellers and borrowers, who are the short side in their respective markets, go to zero, while the measures of buyers and lenders go to positive limits. For each asset, the measure of lenders converges to the asset supply S, and the measure of buyers to a limit m_b. On the other hand, the measures of sellers and borrowers are asymptotically equal to g_s/λ and $g_{\underline{bo}}/\nu$ for two constants g_s and $g_{\underline{bo}}$.

[16] More precisely, we assume that λ and ν go to ∞, holding the ratio $n \equiv \nu/\lambda$ constant. When taking this limit, we will say that a variable Z is asymptotically equal to $z_1/\lambda + z_2/\nu$, if $Z = z_1/\lambda + z_2/(n\lambda) + o(1/\lambda)$.

Proposition 6. *In the symmetric equilibrium of Proposition 5, both assets $i \in \{1,2\}$ have the same price which is asymptotically equal to*

$$p = \frac{\delta + \overline{x} - y}{r} - \frac{\overline{\kappa}}{\lambda m_b}\frac{\overline{x}}{r} - \frac{\phi(r+\overline{\kappa}+2g_s)}{\lambda(1-\phi)m_b}\frac{\overline{x}}{r} + \frac{g_{\underline{bo}}}{r+\overline{\kappa}+\underline{\kappa}\frac{g_s}{r+\overline{\kappa}+\underline{\kappa}+g_s}+g_{\underline{bo}}}\frac{w}{r}, \tag{12}$$

and the same lending fee which is asymptotically equal to

$$w = \theta\left(r+\overline{\kappa}+\underline{\kappa}\frac{g_s}{r+\overline{\kappa}+\underline{\kappa}+g_s}+g_{\underline{bo}}\right)\Sigma, \tag{13}$$

where

$$\Sigma = \frac{\underline{x} - \frac{r+\overline{\kappa}+\underline{\kappa}+g_s}{r+\overline{\kappa}+g_s}(2y-\overline{x})}{2\nu(1-\theta)S}. \tag{14}$$

The price is the sum of four terms. The first term, $(\delta+\overline{x}-y)/r$, is the limit to which the price converges when search frictions go to zero. Not surprisingly, this is the Walrasian price of Propositions 1 and 3. The remaining terms are adjustments to the Walrasian price due to search frictions. The second term is a liquidity discount arising because high-valuation buyers expect to incur a search cost when seeking to unwind their long positions. This cost reduces their valuation and lowers the price. The liquidity discount decreases in the measure of buyers (m_b in the limit) because this reduces the time to sell the asset, and increases in the rate $\overline{\kappa}$ of reversion to average valuation because this reduces the investment horizon. Interpreting the search cost as a transaction cost, the liquidity discount is in the spirit of Amihud and Mendelson (1986).[17]

The third term is a discount arising because high-valuation buyers have bargaining power in the search market and can extract some surplus from the sellers. This "bargaining" discount is present only when the buyers' bargaining power ϕ is non-zero.

The last term is a specialness premium, arising because high-valuation agents can earn a fee by lending the asset in the repo market. As in Proposition 4, lenders do not compete the fee down to zero because the search friction enables them to extract some of the borrowers' short-selling surplus Σ. The fee is an additional cash flow derived from the asset and raises its price. The specialness premium is the PV of the asset's expected lending revenue, but is smaller that the PV w/r of

[17] Consistent with Amihud and Mendelson, the liquidity discount $\overline{\kappa}\overline{x}/(\lambda m_b r)$ is the PV of transaction costs incurred by a sequence of marginal buyers. Indeed, a high-valuation investor (the marginal buyer) reverts to average valuation at rate $\overline{\kappa}$. He then incurs an opportunity cost $\overline{x}$ of holding the asset, since he does not realize the hedging benefit, until he meets a new buyer at rate λm_b.

a continuous stream of the lending fee. This is because lenders must search for borrowers and cannot ensure that their asset is on loan continuously. In fact, the time to meet a borrower does not converge to zero when search frictions become small. For small frictions, the flow of borrowers who enter the market are matched almost instantly with lenders. Because, however, lenders are in positive measure, the meeting time from any given lender's viewpoint is finite.[18]

The short-selling surplus Σ increases in the hedging benefit $\underline{x}$ of the low-valuation agents. It also increases in g_s, the Poisson intensity at which sellers can be contacted in the limit.[19] The easier the sellers are to contact, the more attractive a short-sale becomes to a low-valuation agent because it is easier to buy the asset back.

We next turn to equilibria in which low-valuation agents borrow one asset only. Propositions 7 and 8 establish our main result: there exists an asymmetric equilibrium in which short-selling is concentrated in one asset, and that asset trades at a higher price.

Proposition 7. *Suppose that Equation (11) holds, $\phi \neq 1$, and $\theta \neq 0, 1$. Then, for large λ and ν, there exists an asymmetric equilibrium where short-selling is concentrated in asset 1.*

The mechanism behind the asymmetric equilibrium is explained in the beginning of Section 4: liquidity attracts short-sellers, who in turn bring more liquidity because they generate more volume. The link between liquidity and volume arises because of the search friction that the more buyers and sellers there are, the easier it becomes to locate them. That the presence of an additional trader makes it easier for others to trade is known as search externalities.[20]

In addition to search externalities, our theory relies on the presence of short-sellers and the constraint that these must deliver the same asset they borrowed. Indeed, in the absence of short-sellers, Proposition 2 shows that the assets are equally easy to sell and trade at the same price. The same would hold even with short-sellers if they could deliver any asset and not necessarily the one they borrowed.

In the proof of Proposition 7 we determine the asymptotic behavior of the equilibrium. We show that for each asset i, the measure of lenders converges to the asset supply S, and the measure

[18] Formally, the measure of borrowers is asymptotically equal to $g_{\underline{bo}}/\nu$, and thus the Poisson intensity $\nu\mu_{\underline{bo}}$ at which borrowers can be contacted converges to $g_{\underline{bo}}$.

[19] The Poisson intensity at which sellers can be contacted is $\lambda\mu_s$, and converges to g_s because the measure of sellers is asymptotically equal to g_s/λ. The surplus is increasing in g_s because $2y > \overline{x}$ from Assumption 1.

[20] See, for example, Diamond (1982). Externalities arise in our model because we assume a matching technology with increasing returns to scale: the flow of matches more than doubles when the measures of buyers and sellers double. Increasing returns are plausible in a financial-market context because they imply an increasing volume-liquidity relationship that seems consistent with the empirical evidence.

of buyers to a limit $\hat{m}_{bi}$ such that $\hat{m}_{b1} > \hat{m}_{b2}$. On the other hand, the measures of sellers and borrowers are asymptotically equal to $\hat{g}_{si}/\lambda$ and $\hat{g}_{b\underline{o}}/\nu$ for constants $\hat{g}_{s1} > \hat{g}_{s2}$ and $\hat{g}_{b\underline{o}}$.

Proposition 8. *In the asymmetric equilibrium of Proposition 7, asset prices are asymptotically equal to*

$$p_1 = \frac{\delta + \overline{x} - y}{r} - \frac{\overline{\kappa}}{\lambda \hat{m}_{b1}} \frac{\overline{x}}{r} - \frac{\phi}{\lambda(1-\phi)} \left[\frac{r + \overline{\kappa} + \hat{g}_{s1}}{\hat{m}_{b1}} + \frac{\hat{g}_{s2}}{\hat{m}_{b2}} \right] \frac{\overline{x}}{r} + \frac{\hat{g}_{b\underline{o}}}{r + \overline{\kappa} + \underline{\kappa} \frac{\hat{g}_{s1}}{r + \overline{\kappa} + \underline{\kappa} + \hat{g}_{s1}}} + \hat{g}_{b\underline{o}} \frac{w_1}{r} \quad (15)$$

and

$$p_2 = \frac{\delta + \overline{x} - y}{r} - \frac{\overline{\kappa}}{\lambda \hat{m}_{b2}} \frac{\overline{x}}{r} - \frac{\phi}{\lambda(1-\phi)} \left[\frac{r + \overline{\kappa} + \hat{g}_{s2}}{\hat{m}_{b2}} + \frac{\hat{g}_{s1}}{\hat{m}_{b1}} \right] \frac{\overline{x}}{r}. \quad (16)$$

The lending fee for asset 1 is asymptotically equal to

$$w_1 = \theta \left(r + \overline{\kappa} + \underline{\kappa} \frac{\hat{g}_{s1}}{r + \overline{\kappa} + \underline{\kappa} + \hat{g}_{s1}} + \hat{g}_{b\underline{o}} \right) \Sigma_1, \quad (17)$$

where

$$\Sigma_1 = \frac{\underline{x} - \frac{r + \overline{\kappa} + \underline{\kappa} + \hat{g}_{s1}}{r + \overline{\kappa} + \hat{g}_{s1}} (2y - \overline{x})}{\nu(1-\theta)S}. \quad (18)$$

An immediate consequence of Proposition 8 is that the price of asset 1 exceeds that of asset 2. This is because of three effects working in the same direction. First, the liquidity discount is smaller for asset 1 because this asset has a larger buyer pool, i.e., $\hat{m}_{b1} > \hat{m}_{b2}$. Second, the bargaining discount is smaller for asset 1 because the larger buyer pool implies more outside options for sellers.[21] Finally, asset 1 carries a specialness premium because unlike asset 2, it can be lent to short-sellers.

The results of Propositions 7 and 8 can shed light on several puzzling aspects of the on-the-run phenomenon. At a basic level, they can explain why assets with almost identical payoffs, such as on- and off-the-run bonds, can trade at different prices. Our results can also rationalize the apparent paradox that off-the-run bonds are generally viewed as "scarce" and hard to locate, while at the same time being cheaper than on-the-run bonds. We show that off-the-run bonds are indeed scarce from the viewpoint of short-sellers seeking to buy and deliver them. Because, however,

[21] This logic does not apply to buyers because the marginal buyers are the high-valuation agents who are not limited to the seller pool of a specific asset.

scarcity drives short-sellers away from these bonds, it makes them less liquid and less attractive to marginal buyers who are the agents seeking to establish long positions. Finally, our results have the surprising implication that the trading activity of short-sellers can raise, rather than lower, an asset's price. This is because short-sellers increase both the asset's liquidity and specialness.

We next compare the symmetric and asymmetric equilibria.

Proposition 9. *In the asymmetric equilibrium of Proposition 7:*

(i) There are more buyers and sellers of asset 1 than in the symmetric equilibrium.

(ii) There are fewer buyers and sellers of asset 2 than in the symmetric equilibrium.

(iii) The lending fee of asset 1 is higher than in the symmetric equilibrium.

(iv) The prices of the two assets straddle the symmetric-equilibrium price when $\phi = 0$. For other values of ϕ (e.g., 1/2), both prices can exceed the symmetric-equilibrium price.

Since in the asymmetric equilibrium short-selling is concentrated in asset 1, there are more sellers of this asset than in the symmetric equilibrium. There are also more buyers because of the short-sellers who need to buy the asset back. Conversely, asset 2 attracts fewer buyers and sellers than in the symmetric equilibrium.

The lending fee of asset 1 is higher than in the symmetric equilibrium because of two effects. First, because there are more buyers and sellers of asset 1, a short-sale is easier to execute, and the short-selling surplus is higher. Moreover, lenders of asset 1 are in better position to bargain for this surplus because they do not have to compete with lenders of asset 2.

To explain the price results, we recall that prices differ from the Walrasian benchmark because of a liquidity discount, a bargaining discount, and a specialness premium. In the asymmetric equilibrium, asset 1's liquidity discount is smaller than in the symmetric equilibrium because there are more buyers. Moreover, asset 1's specialness premium is higher because of the higher lending fee. Conversely, asset 2's liquidity discount is higher than in the symmetric equilibrium, and its specialness premium is zero. Therefore, absent the bargaining discount, i.e., when the buyers' bargaining power ϕ is zero, asset 1 trades at a higher price and asset 2 at a lower price relative to the symmetric equilibrium. Quite surprisingly, however, both assets can trade at a higher price because of the bargaining discount. To explain the intuition, we recall that short-sellers exit the seller pool faster when the asset they have borrowed has a larger buyer pool. This occurs in the

asymmetric equilibrium because asset 1 has more buyers than either asset in the symmetric equilibrium. Therefore, there are fewer short-sellers in the asymmetric equilibrium, and the aggregate seller pool can be smaller. This can worsen the buyers' bargaining position and raise the prices of both assets.

4.4 Arbitrage

Since asset prices differ in the asymmetric equilibrium, a natural question is whether there exists a profitable arbitrage. Our analysis so far does not address this question because agents are restricted to hold either long or short positions. In this section we introduce an additional agent group, the "arbitrageurs," who in addition to portfolios allowed to other agents, can hold an arbitrage portfolio that is one share long and one short. We assume that arbitrageurs have average valuation and never switch to high or low. Consistent with the risk-aversion interpretation of utility flows, we set the flow from an arbitrage portfolio to zero. Finally, we assume that arbitrageurs are in infinite measure so that they can hold an unlimited collective position. Proposition 10 shows that the asymmetric equilibrium can be robust to the presence of arbitrageurs.

Proposition 10. *Consider the asymmetric equilibrium of Proposition 7. If $\nu/\lambda \in (n_1, n_2)$ for two positive constants n_1, n_2, then arbitrageurs find it optimal to stay out of the market.*

To explain why arbitrage can be unprofitable, suppose that an arbitrageur attempts to profit from the price differential by buying asset 2 and shorting asset 1. This strategy is unprofitable if

$$p_1 - p_2 < \frac{w_1}{r}, \tag{19}$$

i.e., the price differential does not exceed the PV of asset 1's lending fee.[22] The price differential depends on the lending fee through the specialness premium. Therefore, to check whether Equation (19) holds, we need to substitute the equilibrium values of p_1 and p_2 from Proposition 8:

$$\frac{(\phi r + \overline{\kappa})}{\lambda(1-\phi)}\left[\frac{1}{\hat{m}_{b2}} - \frac{1}{\hat{m}_{b1}}\right]\frac{\overline{x}}{r} + \frac{\hat{g}_{\underline{bo}}}{r + \overline{\kappa} + \underline{\kappa}\frac{\hat{g}_{s1}}{r+\overline{\kappa}+\underline{\kappa}+\hat{g}_{s1}}} + \hat{g}_{\underline{bo}}\frac{w_1}{r} < \frac{w_1}{r}.$$

The first term on the left-hand side reflects asset 1's lower liquidity and bargaining discounts relative to asset 2, and we refer to it as asset 1's liquidity premium. By buying asset 2 and shorting asset

[22] In the proof of Proposition 10 we show that the strategy is unprofitable under the weaker condition $p_1 - p_2 < w_1/r + \xi$, for some transaction cost ξ of establishing the arbitrage position. The cost ξ arises because it is not possible to set up the two legs of the position simultaneously given the Poisson arrival of trading opportunities.

1, an arbitrageur capitalizes on this premium. The arbitrageur also capitalizes on the specialness premium, which is the second term on the left-hand side. Crucially, however, the specialness premium is only a fraction of the cost w_1/r of the arbitrage because lenders cannot ensure that their asset is on loan continuously (as emphasized in Section 4.3). Thus, Equation (19) is satisfied when the lending fee is large enough.[23]

Consider next the opposite strategy of buying asset 1 and shorting asset 2. In the proof of Proposition 10 we show that this strategy is unprofitable if

$$\frac{\hat{g}_{\underline{bo}}}{r + \underline{\kappa}\frac{\hat{g}_{s1}}{r+\underline{\kappa}+g_{s1}} + \hat{g}_{\underline{bo}}}\frac{w_1}{r} \leq p_1 - p_2. \tag{20}$$

The left-hand side is the arbitrageur's fee income from lending asset 1 in the repo market. This exceeds the specialness premium (included in $p_1 - p_2$) because the arbitrageur can hold asset 1 forever, thus being a better lender than a sequence of high-valuation agents. Because, however, the arbitrageur loses on the liquidity premium (the remaining part of $p_1 - p_2$), Equation (20) is satisfied when the lending fee is not to large. In the proof of Proposition 10 we show that Equations (19) and (20) are jointly satisfied when the ratio ν/λ of relative frictions in the spot and the repo market lies in some interval (n_1, n_2). This interval can be quite large as evidenced by the calibration exercise in Section 5.[24]

4.5 Equilibrium Selection

Our model has two identical asymmetric equilibria: one in which short-sellers concentrate in asset 1 and one in which they concentrate in asset 2. The mere existence of multiple equilibria is reminiscent of Boudoukh and Whitelaw (1991) and Boudoukh and Whitelaw (1993) who document that in the

[23] Our analysis has an interesting similarity to Krishnamurthy (2002), who assumes that $p_1 - p_2 = v + zw_1/r$, where v is a "liquidity benefit" of on-the-run bonds, and $z < 1$ is the extent to which bond holders can exploit the specialness premium. In our setting, v is the liquidity premium and z is determined by the lenders' search times.

One might argue that because of same-day settlement in the repo market, some sophisticated investors can manage to lend their asset almost continuously, i.e., $z \approx 1$. We conjecture that in a model with heterogenous lenders, the less sophisticated ones would have a lower reservation value for owning the asset and hence could be the "marginal buyers" in the spot market. The parameter z could then be significantly different than one, reflecting marginal buyers' inferior lending ability.

[24] Equations (19) and (20) ensure that arbitrage portfolios are suboptimal for arbitrageurs, i.e., average-valuation agents with *no* initial position. They do not apply, however, to average-valuation agents with "inherited" positions. Consider, for example, a low-valuation agent with a short position in asset 1, who reverts to average valuation. The agent can unwind the short position by trading with a seller of asset 1, but might also accept to trade with a seller of asset 2. This would hedge the short position, lowering the cost of waiting for a seller of asset 1. In our analysis, we rule out such strategies by assuming that arbitrage portfolios can be held only by arbitrageurs. This is partly for simplicity, to keep agents' life-cycles manageable. One could also argue that many investors do not engage in such strategies because of costs to managing multiple positions, settlement costs, etc. (These costs could be smaller for sophisticated arbitrageurs.) Needless to say, it would be desirable to relax this assumption.

Japanese government-bond market, liquidity concentrates in an arbitrary "benchmark" bond. In the US Treasury market, however, multiple equilibria do not explain why short-sellers concentrate systematically in the on- rather than the off-the-run bond. In this section we explore this issue by considering the case where asset supplies differ. Without loss of generality, we take asset 1 to be in larger supply, i.e., $S_1 > S_2$.

Proposition 11. *As λ and ν become large:*

(i) An equilibrium where short-selling is concentrated in asset 1 exists for all values of $S_1 - S_2$.

(ii) An equilibrium where short-selling is concentrated in asset 2 exists for a set of values of $S_1 - S_2$ that converges to $[0, \hat{S}]$ with $\hat{S} > 0$.

(iii) An equilibrium where low-valuation agents short-sell both assets exists for a set of values of $S_1 - S_2$ that converges to $\{0\}$.

Proposition 11 shows that asset supply is a natural device in selecting among equilibria. For small search frictions, the symmetric equilibrium ceases to exist as long as asset supplies differ. Moreover, if the difference exceeds $\hat{S}$, the equilibrium in which short-sellers concentrate in asset 2 ceases to exist as well. The only remaining equilibrium is that short-sellers concentrate in asset 1. Intuitively, short-sellers prefer the asset with the larger seller pool because they can buy it back more easily. Since asset 1 is in larger supply, it has more owners who eventually turn into sellers. If the supply difference $S_1 - S_2$ is large enough, this effect makes the seller pool of asset 1 larger than the seller pool of asset 2, even if all short-sellers were to concentrate in asset 2. Therefore, if $S_1 - S_2$ is large enough, the equilibrium in which short-sellers concentrate in asset 2 does not exist.

Proposition 11 is a first step towards reconciling our theory based on multiple equilibria with the empirical fact that liquidity in the US Treasury market concentrates systematically in just-issued bonds. Indeed, a commonly-held view is that a bond's effective supply decreases over time as the bond becomes "locked away" in the portfolios of buy-and-hold investors (see Amihud and Mendelson (1991)). Our theory cannot explain the decrease in effective supply since there are no auction-cycle dynamics. Proposition 11 suggests, however, that if off-the-run bonds are indeed in smaller effective supply, they are less likely to attract short-sellers and for that reason less liquid.

4.6 Comparative Statics

In this section we explore the comparative statics of our model, and draw connections with empirical studies. We examine how liquidity, specialness, and the price premium depend on the extent of short-selling activity and on assets' supplies (i.e., issue sizes). We measure asset i's specialness by the ratio w_i/p_i of the lending fee to the price, the price premium by $p_1 - p_2$, and the short-selling activity by the flow $\underline{F}$ of short-sellers entering the market. Liquidity can be measured by search times, but these can differ for buyers and sellers. To condense search times into an one-dimensional measure, we multiply the expected search time for buying asset i with that for selling the asset, and take the inverse. This measure has the advantage of being equal, up to the multiplicative constant λ, to asset i's trading volume, i.e., the flow of matches $\lambda\mu_{bi}\mu_{si}$ between buyers and sellers.[25]

Proposition 12. *Consider the asymmetric equilibrium in which short-selling is concentrated in asset 1, and suppose that λ and ν are large.*

(i) If the flow $\underline{F}$ of short-sellers increases, then asset 1's liquidity increases, asset 2's liquidity stays constant, asset 1's specialness increases, and the price premium increases.

(ii) If the supply S_1 of asset 1 decreases, holding asset 2's supply S_2 constant, then asset 1's liquidity decreases, asset 2's liquidity stays constant, asset 1's specialness can increase or decrease, and the price premium can increase or decrease.

Result (i) is one of our theory's main implications: short-selling activity drives both the superior liquidity of on-the-run bonds, and their specialness. By concentrating in asset 1, short-sellers increase that asset's trading volume and liquidity. They also generate specialness because they demand asset 1 in the repo market. Liquidity and specialness raise asset 1's price above that of asset 2. These results are consistent with several empirical studies. Jordan and Jordan (1997) provide a case study where short-seller demand for a particular Treasury note generated a large price premium. Krishnamurthy (2002) measures short-seller demand by the issuance of corporate and agency bonds, arguing that dealers short Treasuries to hedge their inventories. He finds that issuance is positively related to the on-the-run premium. Graveline and McBrady (2004) emphasize the role of short-seller demand using a conceptual framework very similar to ours. They construct several measures of demand, attempting to get both at the hedging and the speculative component. They find that short-seller demand is the strongest determinant of specialness once variation related

[25] An agent's expected search time is the inverse of the Poisson intensity of arrival of counterparties. Thus, for a buyer of asset i it is $1/(\lambda\mu_{si})$, and for a seller it is $1/(\lambda\mu_{bi})$.

to the auction cycle is taken out. Moulton (2004) also finds evidence linking short-selling demand to specialness.

Result (ii) shows that a decrease in asset 1's supply decreases liquidity, but can increase or decrease specialness. Specialness can increase because of a scarcity effect in the repo market: since there are fewer lenders of asset 1, they can extract a higher fee from short-sellers. There is, however, an offsetting scarcity effect in the spot market: because there are fewer sellers of asset 1, the asset is harder to deliver. This reduces short-sellers' willingness to borrow the asset, and can reduce the lending fee that lenders are able to extract. Furthermore, if supply drops below a threshold, an equilibrium with short-seller concentration in asset 1 is not possible, as shown in Proposition 11. Short-sellers migrate to asset 2, and asset 1's specialness drops discontinuously.

Result (ii) can help interpret the variation of liquidity and specialness over the auction cycle. Graveline and McBrady (2004) document that specialness increases while a bond is on-the-run, but drops discontinuously with the issuance of the new bond. At the same time, Fleming (2002) documents that liquidity decreases steadily with time from issuance. Suppose for reasons outside our model, that a bond's effective supply decreases over time as the bond becomes "locked away" in the portfolios of buy-and-hold investors. Then the bond's liquidity decreases, but specialness can increase as the bond becomes scarcer in the repo market. Eventually, however, specialness drops because short-sellers migrate to the new bond, which is easier to buy in the spot market.[26]

More broadly, Result (ii) can reconcile our model with the empirical fact that liquidity and specialness do not always move in the same direction. For example, a sudden decrease in a bond's effective supply (such as a short squeeze) can lead to a jump up in specialness but a drop in liquidity. Of course, our model cannot fully address such phenomena because it is stationary. A comprehensive time-series analysis would require modelling auction-cycle dynamics and/or stochastic shocks.

5 Calibration

In this section we perform a calibration exercise, and show that our model can generate significant price effects even for short search times. For the calibration we extend the model to more than two assets. This provides a more accurate description of the US Treasury market, where there is

[26] Empirical studies generally document a decreasing relationship between supply and specialness or on-the-run premia. Cornell and Shapiro (1989) and Jordan and Jordan (1997) provide case studies where large price premia were generated by a short-squeeze or a large investor's unwillingness to lend, respectively. Krishnamurthy (2002) finds that on-the-run premia are negatively related to issue size, and Graveline and McBrady (2004) and Moulton (2004) find a negative relationship between issue size and specialness.

one on-the-run and multiple off-the-run securities for each maturity range. With multiple assets there is again an equilibrium in which short-sellers concentrate in one asset, e.g., asset 1. To compute this equilibrium for the purpose of calibration, we do not rely on the asymptotic closed-form solutions of Section 4.3. Instead, we use a simple numerical algorithm that solves the exact system of equations and checks that arbitrage is unprofitable. Table 1 lists our chosen values for the exogenous parameters, and Tables 2 and 3 list the calibration results.

We set the number of assets to $I = 20$, consistent with the fact that on-the-run bonds account for about 5% of the Treasury market capitalization (Dupont and Sack (1999)). We assume that all assets are in identical supply S. We normalize the total supply IS to one, without loss of generality: Equations (1)-(5) and (B.3)-(B.7) show that if $(S, \overline{F}, \underline{F}, 1/\lambda, 1/\nu)$ are scaled by the same factor, the meeting intensities of each investor type stay the same.

As in the case of two assets, we assume that demand exceeds supply, i.e., $\overline{F}/\overline{\kappa} > IS + \underline{F}/\underline{\kappa}$. We select $(\overline{F}, \underline{F})$ to make this an approximate equality; otherwise for small frictions, search times for sellers would be much shorter than for buyers. We use the second degree of freedom in $(\overline{F}, \underline{F})$ to match the level of short-selling activity. Namely, in our calibration the amount of ongoing repo agreements for asset 1 is about seven times the asset's issue size (Table 2), which is within reasonable range.[27]

The expected investment horizons $1/\overline{\kappa}$ and $1/\underline{\kappa}$ are chosen to match turnover. Sundaresan (2002) and Strebulaev (2002) report that on-the-run bonds trade about ten times more than their off-the-run counterparts. Since the entire stock of Treasury securities turns over in less than three weeks (Dupont and Sack (1999)), on-the-run bonds turn over in about two-thirds of a day, and off-the-run bonds in about 125 days.[28] In our model the turnover of off-the-run bonds is generated by high-valuation investors. We let $1/\overline{\kappa} = 0.5$ years, i.e., 125 trading days, implying a turnover time of about the same (Table 2). The turnover of on-the-run bonds is generated mainly by short-sellers. We let $1/\underline{\kappa} = 0.025$ years, i.e., about six trading days. Such a short horizon could be reasonable for dealers in corporate bonds or mortgage-backed securities who have transitory needs to hedge inventory. For our chosen value of $\underline{\kappa}$, asset 1 turns over in 0.88 days, and its volume relative to

[27] For example, on February 2, 2005, primary dealers reported asset loans of about $2 trillion (New York Fed website, `www.ny.frb.org/markets/gsds/search.cfm`). Since the Treasury market is worth about $4 trillion, of which 5% are on-the-run bonds, the amount of repo agreements exceeds the market value of on-the-run bonds by about $2/(4 \times 5\%) = 10$. We select a number below ten to account for repo activity in off-the-run bonds. A higher number would strengthen our results because the lending fee would increase.

[28] Suppose, for example, that the average Treasury security turns over in twelve trading days. Since on-the-run bonds account for about 5% of market capitalization and 10/11 of trading volume, they turn over in $5\% \times 12/(10/11) = 0.66$ days, while off-the-run bonds turn over in $95\% \times 12/(1/11) = 125.4$ days.

the aggregate of the other assets is 7.5 (Table 2).[29] This is lower than the actual value of ten, but one could argue that short-selling is not the only factor driving the large relative volume of on-the-run bonds. Furthermore, raising the relative volume by increasing $\underline{\kappa}$ would strengthen our results because the lending fee would increase.

The parameters λ and ν are chosen to generate short search times, as reported in Table 2. Assuming ten trading hours per day,[30] most search times are in the order of a few hours or less, significantly smaller than the standard settlement time. It takes 12 minutes to sell the "on-the-run" asset 1 and 2.7 hours to buy it. Each "off-the-run" asset $i \in \{2, .., I\}$ can be sold in 2.8 hours and bought in 2.2 days. The time to buy might seem long, but is not unreasonable given that locating a specific off-the-run issue is often viewed as difficult. Furthermore, in our model all off-the-run issues are perfect substitutes for their buyers, who are the high-valuation agents. Therefore, a buyer's effective search time does not exceed $2.2/(I-1) = 0.11$ days. Finally, it takes 42 minutes to borrow asset 1 in the repo market and 8.7 hours to lend it. The time to lend the on-the-run asset might seem long but could be interpreted as an average across asset owners, some of whom do not engage in asset lending in practice.

Table 1: Parameter Values used in the Numerical Example.

Parameters		Value
Number of assets	I	20
Supply of each asset	S	0.05
Flow of high-valuation investors	$\overline{F}$	2.7
Flow of low-valuation investors	$\underline{F}$	13.6
Switching intensity of high-valuation investors	$\overline{\kappa}$	2
Switching intensity of low-valuation investors	$\underline{\kappa}$	40
Contact intensity in spot market	λ	10^6
Contact intensity in repo market	ν	7.5×10^4
Bargaining power of a buyer	ϕ	0.5
Bargaining power of a lender	θ	0.5
Riskless rate	r	4%
Dividend rate	δ	1
Hedging benefit of high-valuation investors	$\overline{x}$	0.4
Cost of risk bearing	y	0.5

The parameters ϕ and θ are set to 0.5 so that all agents are symmetric. The riskless rate r is set

[29] The six-day expected horizon of short-sellers is approximately equal to the turnover time of the asset supply that they generate ($\underline{F}/\underline{\kappa}$). This supply is about seven times the issue size S, and turns over seven times more slowly.

[30] US Treasury securities are traded round the clock in New York, London, and Tokyo. However, Fleming (1997) reports that 94% of the trading takes place in New York from 7:30am to 5:30pm.

to 4%, consistent with Ibbotson (2004)'s average T-bill rate of 3.8% during the period 1926-2002. Given that prices and lending fees are linear in $(\delta, \overline{x}, \underline{x}, y)$, we let $\delta = 1$ and report relative prices (e.g., δ/p, w/p). The parameters $\overline{x}$ and y are selected based on assets' risk premia, measured by the difference $\delta/p_i - r$ between expected returns and the riskless rate. We assume that $\overline{x} < y$, so that the Walrasian price $(\delta + \overline{x} - y)/r$ incorporates a positive premium. We also assume that $y < \delta$, so that risk premia do not result in negative prices: the lowest possible price is $(\delta - y)/r$, the PV of the average-valuation agents' certainty equivalent. Our chosen values of $\overline{x}$ and y generate risk premia of about 2-2.5%, which are within reasonable range for government bonds. (For example, Ibbotson (2004) reports that long-term Treasuries returned 1.9% per year above bills during the period 1926-2002.[31]) To select $\underline{x}$, we show in Appendix E that the CARA-based foundation of our model suggests the restriction $\underline{x} \leq 4y - \overline{x} = 1.6$ because otherwise low-valuation agents would prefer to short more than one share. Moreover, our numerical calculations indicate that $\underline{x}$ must exceed 0.97 so that the lending fee is large enough to preclude arbitrage. We therefore assume $0.97 \leq \underline{x} \leq 1.6$, and report results for the two extreme values.

Table 2: Numerical Results: Search Times and Turnover.

Variable		Value
Average time to sell asset 1	$1/(\lambda\mu_{b1})$	0.02 days
Average time to buy asset 1	$1/(\lambda\mu_{s1})$	0.27 days
Average time to sell asset $i \in \{2, .., I\}$	$1/(\lambda\mu_{bi})$	0.28 days
Average time to buy asset $i \in \{2, .., I\}$	$1/(\lambda\mu_{si})$	2.20 days
Average time to borrow asset 1	$1/(\lambda\mu_{\overline{\ell}1})$	0.07 days
Average time to lend asset 1	$1/(\lambda\mu_{\underline{bo}})$	0.87 days
Time to turn over stock of asset 1	$S/(\lambda\mu_{b1}\mu_{s1})$	0.88 days
Time to turn over stock of asset $i \in \{2, .., I\}$	$S/(\lambda\mu_{bi}\mu_{si})$	125.28 days
Volume of asset 1 vs. aggregate of assets $i \in \{2, .., I\}$	$(\lambda\mu_{b1}\mu_{s1})/((I-1)\lambda\mu_{bi}\mu_{si})$	7.50
Repo agreements for asset 1 relative to issue size	$\mu_{\overline{n}1}/S$	7.03

Table 3 reports the prices and lending fees.[32] When $\underline{x}$ is equal to its lowest value of 0.97, the effects are quite small: assets' expected returns differ by 4bp, and specialness is 3bp. When, however, $\underline{x}$ is equal to its highest value of 1.6, the effects are large and consistent with empirical findings. In particular, the 51bp difference in expected returns is consistent with Warga (1992), who

[31] Of course, this is only suggestive since in our model risk arises because of asset payoffs and not interest rates.

[32] A feature of Table 3 that might appear surprising is that the prices $p_1 = 1/6.48\% = 15.43$ and $p_2 = 1/6.52\% = 15.38$ differ significantly from the Walrasian limit $(\delta + \overline{x} - y)/r = 22.5$. This is because demand and supply are very close: when they are exactly equal, any price between $(\delta - y)/r = 15$ and 22.5 is consistent with Walrasian equilibrium.

reports that on-the-run portfolios return 55bp below matched off-the-run portfolios.[33] Moreover, the lending fee of 35bp is consistent with Duffie (1996), who reports a specialness difference of 40bp between on- and off-the-run bonds.[34]

Table 3: Numerical Results: Prices and Lending Fees.

Variable		$\underline{x} = 0.97$	$\underline{x} = 1.6$
Expected return of asset 1	δ/p_1	6.48%	6.02%
Expected return of asset $i \in \{2, .., I\}$	δ/p_i	6.52%	6.53%
Spread	$\delta/p_i - \delta/p_1$	4bp	51bp
Lending fee	w_1/p_1	3bp	35bp

The large price effects are in spite of the short search times. The transaction costs implicit in these times are, in fact, very small. For example, the cost incurred by a high-valuation buyer is not to receive the hedging benefit $\overline{x}$ while searching. With a search time not exceeding 0.11 days, i.e., 0.11/250 of a year, the search cost is a fraction $\overline{x} \times (0.11/250) \times (1/6.53\%) = 1.1 \times 10^{-5}$ of the price, i.e., 0.11 cents per \$100 transaction value. Likewise, the search cost of a low-valuation agent seeking to borrow asset 1 in the repo market is not to receive the hedging benefit $\underline{x}$. When $\underline{x}$ is equal to its highest value of 1.6, the cost is a fraction $\underline{x} \times (0.07/250) \times (1/6.53\%) = 2.9 \times 10^{-5}$ of the price, i.e., 0.29 cents per \$100 transaction value. Such costs are smaller than the average bid-ask spread in the Treasury market, which is 1.1 cent (Dupont and Sack (1999)). While this raises the question of what drives the bid-ask spread, it also shows that the large price effects in our model are driven by very small transaction costs.[35]

Small transaction costs imply that most of the return spread is due to the specialness premium. Generalizing the decomposition in Section 4.3, we can show that when $\underline{x} = 1.6$ the specialness premium accounts for 99% of the spread while the liquidity premium for only 1%. Of course, this

[33]Some studies find smaller effects. For example, Goldreich, Hanke, and Nath (2002) report that on-the-run bonds yield 1.5bp below off-the-run bonds, and Fleming (2003) reports 5.6bp. These papers, however, focus on bonds with a long time to maturity, for which the three-month convenience yield of being on-the-run has only a small effect on the yield to maturity. Warga (1992) compares the returns of on- and off-the-run bond portfolios rather than their yields to maturity. This isolates the on-the-run convenience yield in exactly the same way as in this paper. Amihud and Mendelson (1991) compare yields to maturity, but can isolate the convenience yield because they focus on securities with very short times to maturity. They find that Treasury bills maturing in less than six months yield 38bp below comparable Treasury notes.

[34]The expected return spread $\delta/p_i - \delta/p_1$ in Table 3 is greater than the lending fee w_1/p_1. This suggests an arbitrage strategy of shorting \$1 of asset 1, paying the lending fee, and buying \$1 of asset 2. The payoff of this strategy is risky, however, because the assets are held in different quantities. Adjusting for risk amounts to calculating the marginal utility flow $(\delta - y)/p_i - (\delta - y)/p_1 - w_1/p_1$ that an arbitrageur would derive, which turns out to be negative.

[35]Our model could generate larger transaction costs if agents had to incur a monetary cost of search (e.g., pay a broker), in addition to not receiving hedging benefits.

does not mean that liquidity does not matter; it rather means that liquidity can have large effects because it induces short-seller concentration and creates specialness.

6 Conclusion

This paper proposes a search-based theory of the on-the-run phenomenon. We argue that liquidity and specialness are not independent explanations of this phenomenon, but can be explained simultaneously by short-selling activity. Short-sellers in our model can endogenously concentrate in one of two identical assets, because of search externalities and the constraint that they must deliver the asset they borrowed. That asset enjoys greater liquidity, measured by search times, and a higher lending fee ("specialness"). Moreover, liquidity and specialness translate into price premia which are consistent with no-arbitrage. We derive closed-form solutions in the realistic case of small frictions, and show that a calibration can generate effects of the observed magnitude.

While our analysis is motivated from the government-bond market, some lessons can be more general. Perhaps the main lesson concerns the law of one price - a fundamental tenet of Finance. We show that this law can be violated in a significant manner in a model where all agents are rational but the trading mechanism is not Walrasian. Our search-based trading mechanism is of course an idealization, but it captures the bilateral nature of trading in over-the-counter markets. Furthermore, the search times that are needed to generate significant price differentials are small, in the order of a few hours. For such times, it is unclear whether the search framework is a worse description of over-the-counter markets than a Walrasian auction, which assumes multilateral trading.

A Proofs of Propositions 1-4

Proof of Proposition 1: At time t, an agent with valuation x_t chooses an asset i and a position q in the asset to solve

$$\max_{i\in\{1,2\}} \max_{q\in\{0,1\}} \left[q(\delta + x_t) - |q|\, y - qrp_i\right], \tag{A.1}$$

i.e., maximize the flow utility minus the time value of the position's cost. In equilibrium, assets trade at the same price because otherwise no agent would demand a long position in the more expensive asset. Denoting by p the common price, no agent would demand a long position in any asset if $rp > (\delta + \overline{x} - y)$. Conversely, if $rp < (\delta + \overline{x} - y)$, then high-valuation agents would demand long positions, which generates excess demand from Assumption 2. Therefore, $rp = (\delta + \overline{x} - y)$. Under this price, high-valuation agents are indifferent between a long and no position, and all other agents hold no position. ∎

Proof of Proposition 2: In equilibrium, either high-valuation agents accept to buy asset i, or they refuse to do so and the asset is owned only by average-valuation agents. To nest the two cases, we define the variable λ_i by $\lambda_i \equiv \lambda$ if high-valuation agents accept to buy asset i and $\lambda_i \equiv 0$ otherwise. We denote by $V_{\overline{b}}$ the utility of a high-valuation agent not owning an asset, $V_{\overline{n}i}$ the utility of a high-valuation agent owning asset i, and $V_{\overline{s}i}$ the utility of an average-valuation agent owning asset i. These utilities satisfy the equations

$$rV_{\overline{b}} = -\overline{\kappa} V_{\overline{b}} + \sum_{i=1}^{2} \lambda_i \mu_{\overline{s}i} \left(V_{\overline{n}i} - p_i - V_{\overline{b}}\right), \tag{A.2}$$

$$rV_{\overline{n}i} = \delta + \overline{x} - y + \overline{\kappa}\left(V_{\overline{s}i} - V_{\overline{n}i}\right), \tag{A.3}$$

$$rV_{\overline{s}i} = \delta - y + \lambda_i \mu_{\overline{b}} \left(p_i - V_{\overline{s}i}\right). \tag{A.4}$$

The logic behind these equations is explained in Section 4.2: for example, Equation (A.2) generalizes (6) to the case when a high-valuation agent can refuse to buy an asset. The price of asset i is determined by

$$p_i = \phi V_{\overline{s}i} + (1-\phi)\left(V_{\overline{n}i} - V_{\overline{b}}\right), \tag{A.5}$$

which is Equation (9). Moreover, equilibrium imposes that

$$\lambda_i = \lambda \Leftrightarrow \hat{\Sigma}_i \equiv V_{\overline{n}i} - V_{\overline{b}} - V_{\overline{s}i} \geq 0, \tag{A.6}$$

i.e., high-valuation agents accept to buy asset i if this transaction generates a positive surplus $\hat{\Sigma}_i$.

Subtracting (A.2) and (A.4) from (A.3), and replacing p_i by (A.5), we find

$$(r+\overline{\kappa})\hat{\Sigma}_i = \overline{x} - \phi\sum_{j=1}^{2}\lambda_j\mu_{\overline{s}j}\hat{\Sigma}_j - (1-\phi)\lambda_i\mu_{\overline{b}}\hat{\Sigma}_i. \tag{A.7}$$

If $\lambda_1 = \lambda_2 = 0$, Equation (A.7) implies that $\hat{\Sigma}_i = \overline{x}/(r+\overline{\kappa}) > 0$, a contradiction. If $\lambda_1 = \lambda$ and $\lambda_2 = 0$, Equation (A.7) implies that $\hat{\Sigma}_2 > \hat{\Sigma}_1 > 0$, again a contradiction. Therefore, the only possibility is that $\lambda_1 = \lambda_2 = \lambda$, i.e., high-valuation agents accept to buy both assets. For $\lambda_1 = \lambda_2 = \lambda$, the variables $(V_{\overline{n}i}, V_{\underline{n}i}, p_i, \hat{\Sigma}_i)$ are independent of i, and thus the Law of One Price holds. ∎

Proof of Proposition 3: The lending fee is zero by the argument preceding the proposition's statement. Agents' optimization problem is (A.1) with the only difference that $q \in \{-1, 0, 1\}$. Same arguments as in Proposition 1 imply that assets trade at the same price p, such that $rp \leq (\delta + \overline{x} - y)$. If $rp < (\delta + \overline{x} - y)$, then high-valuation agents would demand long positions, and average-valuation agents would not demand short positions from Assumption 1. This implies excess demand from Assumption 2, and thus $rp = (\delta + \overline{x} - y)$. Under this price, high-valuation agents are indifferent between a long and no position. Moreover, Assumption 1 implies that low-valuation agents hold short positions and average-valuation agents hold no position. ∎

Proof of Proposition 4: If in equilibrium low-valuation agents refuse to borrow asset i, the asset carries no lending fee, and its owners are high-valuation agents who sell when they switch to average valuation. If instead low-valuation agents accept to borrow asset i, some owners can be average-valuation. Indeed, because the asset carries a positive lending fee, its owners might prefer not to terminate a repo contract when they switch to average valuation, but wait until the borrower wishes to terminate. To nest the two cases, we define the variable ν_i by $\nu_i \equiv \nu$ if low-valuation agents accept to borrow asset i and $\nu_i \equiv 0$ otherwise. We denote by $V_{\overline{l}i}$ the utility of a high-valuation agent seeking to lend asset i, $V_{\overline{n}i}$ the utility of a high-valuation agent who is in a repo contract lending asset i, V_{ni} the utility of an average-valuation agent who is in the same repo contract and waits for the borrower to terminate, $V_{\underline{bo}}$ the utility of a low-valuation agent seeking to borrow an asset, and $V_{\underline{n}i}$ the utility of a low-valuation agent who is in a repo contract borrowing

asset i. These utilities satisfy the equations

$$rV_{\overline{\ell}i} = \delta + \overline{x} - y + \overline{\kappa}\left(p_i - V_{\overline{\ell}i}\right) + \nu_i \mu_{\underline{bo}}\left(V_{\overline{n}i} - V_{\overline{\ell}i}\right), \tag{A.8}$$

$$rV_{ni} = \delta - y + w_i + \underline{\kappa}\left(p_i - V_{ni}\right), \tag{A.9}$$

$$rV_{\underline{bo}} = -\underline{\kappa} V_{\underline{bo}} + \sum_{i=1}^{2} \nu_i \mu_{\overline{\ell}i}\left(V_{\underline{n}i} + p_i - V_{\underline{bo}}\right). \tag{A.10}$$

The remaining two equations depend on whether an owner terminates a repo contract immediately upon switching to average valuation, or whether he waits for the borrower to terminate. The condition for immediate termination is $p_i \geq V_{ni}$, and the equations in that case are

$$rV_{\overline{n}i} = \delta + \overline{x} - y + w_i + \overline{\kappa}\left(p_i - V_{\overline{n}i}\right) + \underline{\kappa}\left(V_{\overline{\ell}i} - V_{\overline{n}i}\right), \tag{A.11}$$

$$rV_{\underline{n}i} = -\delta + \underline{x} - y - w_i + \overline{\kappa}\left(V_{\underline{bo}} - p_i - V_{\underline{n}i}\right) + \underline{\kappa}\left(-p_i - V_{\underline{n}i}\right). \tag{A.12}$$

Equation (A.9) implies that $p_i \geq V_{ni}$ is equivalent to $\delta - y + w_i - rp_i \leq 0$. The latter condition is satisfied for small search frictions since $w_i \approx 0$ and $rp_i \approx \delta + \overline{x} - y$. For brevity, we focus on the case $p_i \geq V_{ni}$ from now on, and treat the general case in Appendix E.

To determine the price p_i, note that if $p_i > V_{\overline{\ell}i}$, then high-valuation agents would not demand long positions, and neither would other agents with lower valuations. Conversely, if $p_i < V_{\overline{\ell}i}$, then high-valuation agents would demand long positions. Since the measure of short-sellers does not exceed that of low-valuation agents (and is, in fact, strictly smaller because of the search friction), Assumption 2 implies excess demand for asset i. Therefore, $p_i = V_{\overline{\ell}i}$. The lending fee w_i is such that the lender receives a fraction $\theta \in [0,1]$ of the surplus Σ_i in a repo transaction. It is implicitly determined by

$$V_{\overline{n}i} - V_{\overline{\ell}i} = \theta \Sigma_i, \tag{A.13}$$

where $\Sigma_i \equiv V_{\overline{n}i} - V_{\overline{\ell}i} + p_i + V_{\underline{n}i} - V_{\underline{bo}}$. Finally, equilibrium imposes Equation (10), i.e., low-valuation agents accept to borrow asset i if this transaction generates a positive surplus Σ_i.

Since $p_i = V_{\overline{\ell}i}$, the surplus is $\Sigma_i = V_{\overline{n}i} + V_{\underline{n}i} - V_{\underline{bo}}$. Subtracting Equations (A.10) and (A.12) from (A.11), and noting that Equation (A.13) implies $V_{\underline{n}i} - V_{\underline{bo}} = (1-\theta)\Sigma_i$, we find:

$$(r + \overline{\kappa} + \underline{\kappa})\Sigma_i = \overline{x} + \underline{x} - 2y - (1-\theta)\sum_{j=1}^{2} \nu_j \mu_{\overline{\ell}j} \Sigma_j. \tag{A.14}$$

Equation (A.14) implies $\Sigma_1 = \Sigma_2 \equiv \Sigma$ and thus $\nu_1 = \nu_2$. If $\nu_1 = \nu_2 = 0$, then $\Sigma = (\overline{x} + \underline{x} - 2y)/(r + \overline{\kappa} + \underline{\kappa})$, which is positive by Assumption 1, a contradiction. Therefore, $\nu_1 = \nu_2 = \nu$, i.e., low-valuation agents accept to borrow both assets. For $\nu_1 = \nu_2 = \nu$, the variables $(V_{\overline{\ell}i}, V_{\overline{n}i}, V_{\underline{n}i}, p_i, w_i)$ are independent of i, and thus the Law of One Price holds. ∎

B Population Measures

To compute population measures, it is not necessary to consider the separate types $(\overline{n}\underline{s}i, \overline{n}\underline{n}i, \overline{n}\underline{b}i)$, and we merge them into a type of $\overline{n}i$ of high-valuation non-searchers. We denote this type's measure by $\mu_{\overline{n}i}$, as in Equation (4). We also denote by f_i the inflow from type $\overline{n}i$ to type $\overline{\ell}i$. The inflow-outflow equations are

$$\textit{Lenders } \overline{\ell}i \qquad \lambda\mu_{\overline{b}}\mu_{si} + f_i = \overline{\kappa}\mu_{\overline{\ell}i} + \nu_i\mu_{\underline{bo}}\mu_{\overline{\ell}i} \tag{B.1}$$

$$\textit{Non-searchers } \overline{n}i \qquad \nu_i\mu_{\underline{bo}}\mu_{\overline{\ell}i} = f_i + \overline{\kappa}\mu_{\overline{n}i} \tag{B.2}$$

$$\textit{Sellers } \overline{s}i \qquad \overline{\kappa}\mu_{\overline{\ell}i} + \overline{\kappa}\mu_{\underline{s}i} = \lambda\mu_{bi}\mu_{\overline{s}i} \tag{B.3}$$

$$\textit{Borrowers } \underline{bo} \qquad \underline{F} + \sum_{i=1}^{2} \overline{\kappa}(\mu_{\underline{s}i} + \mu_{\underline{n}i}) = \underline{\kappa}\mu_{\underline{bo}} + \sum_{i=1}^{2} \nu_i\mu_{\underline{bo}}\mu_{\overline{\ell}i} \tag{B.4}$$

$$\textit{Sellers } \underline{s}i \qquad \nu_i\mu_{\underline{bo}}\mu_{\overline{\ell}i} = \overline{\kappa}\mu_{\underline{s}i} + \underline{\kappa}\mu_{\underline{s}i} + \lambda\mu_{bi}\mu_{\underline{s}i} \tag{B.5}$$

$$\textit{Non-searchers } \underline{n}i \qquad \lambda\mu_{bi}\mu_{\underline{s}i} = \overline{\kappa}\mu_{\underline{n}i} + \underline{\kappa}\mu_{\underline{n}i} \tag{B.6}$$

$$\textit{Buyers } \underline{b}i \qquad \underline{\kappa}\mu_{\underline{n}i} = \overline{\kappa}\mu_{\underline{b}i} + \lambda\mu_{\underline{b}i}\mu_{si} \tag{B.7}$$

We determine population measures by the system of (1)-(5) and (B.3)-(B.7). The total number of equations is 18 (because some are for each asset), and the 18 unknowns are the measures of the 14 types $\overline{b}$, $\underline{bo}$, $\{\overline{\ell}i, \overline{n}i, \overline{s}i, \underline{s}i, \underline{n}i, \underline{b}i\}_{i\in\{1,2\}}$ and $\{\mu_{bi}, \mu_{si}\}_{i\in\{1,2\}}$. A solution to the system satisfies Equations (B.1) and (B.2), which is why we do not include them into the system. Indeed, adding Equations (B.5)-(B.7), and using Equation (4), we find

$$\nu_i\mu_{\underline{bo}}\mu_{\overline{\ell}i} = \overline{\kappa}\mu_{\underline{s}i} + \underline{\kappa}\mu_{\overline{n}i} + \lambda\mu_{\underline{b}i}\mu_{si}.$$

Therefore, Equation (B.2) holds with $f_i = \underline{\kappa}\mu_{\underline{s}i} + \lambda\mu_{\underline{b}i}\mu_{si}$. For this value of f_i, Equation (B.1) becomes $\lambda\mu_{bi}\mu_{si} + \underline{\kappa}\mu_{\underline{s}i} = \overline{\kappa}\mu_{\overline{\ell}i} + \nu_i\mu_{\underline{bo}}\mu_{\overline{\ell}i}$, and is redundant because it can be derived by adding Equations (B.3) and (B.5).

To solve the system, we reduce it to a simpler one in the six unknowns $\mu_{\underline{bo}}$, $\mu_{\overline{b}}$, and $\{\mu_{bi}, \mu_{si}\}_{i\in\{1,2\}}$. Adding Equations (B.5) and (B.6), we find

$$\mu_{\underline{s}i} + \mu_{\underline{n}i} = \frac{\nu_i \mu_{\underline{bo}} \mu_{\overline{\ell}i}}{\overline{\kappa} + \underline{\kappa}}. \tag{B.8}$$

Plugging into equation (B.4), and using Equation (3), we find

$$\underline{F} = \underline{\kappa}\mu_{\underline{bo}} + \frac{\underline{\kappa}}{\overline{\kappa} + \underline{\kappa}} \sum_{i=1}^{2} \nu_i \mu_{\underline{bo}} (S - \mu_{si}). \tag{B.9}$$

Now Equations (B.5) and (3) imply that

$$\mu_{\underline{s}i} = \frac{\nu_i \mu_{\underline{bo}} (S - \mu_{si})}{\overline{\kappa} + \underline{\kappa} + \lambda\mu_{bi}}. \tag{B.10}$$

Equation (B.6) implies that

$$\mu_{\underline{n}i} = \frac{\lambda \mu_{\underline{s}i} \mu_{bi}}{\overline{\kappa} + \underline{\kappa}} \tag{B.11}$$

and Equation (B.7) implies that

$$\mu_{\underline{b}i} = \frac{\underline{\kappa}\mu_{\underline{n}i}}{\overline{\kappa} + \lambda\mu_{si}}. \tag{B.12}$$

Combining these equations to compute $\mu_{\underline{b}i}$, and using Equation (1), we find

$$\mu_{bi} = \mu_{\overline{b}} + \frac{\underline{\kappa}\lambda\mu_{bi}\nu_i\mu_{\underline{bo}}(S - \mu_{si})}{(\overline{\kappa} + \underline{\kappa})(\overline{\kappa} + \underline{\kappa} + \lambda\mu_{bi})(\overline{\kappa} + \lambda\mu_{si})}. \tag{B.13}$$

Noting that $\mu_{\overline{\ell}i} + \mu_{\underline{s}i} = S - \mu_{\overline{s}i}$, we can use Equation (B.3) to compute $\mu_{\overline{s}i}$:

$$\mu_{\overline{s}i} = \frac{\overline{\kappa}S}{\overline{\kappa} + \lambda\mu_{bi}}. \tag{B.14}$$

Adding Equations (B.10) and (B.14), and using Equation (2), we find

$$\mu_{si} = \frac{\overline{\kappa}S}{\overline{\kappa} + \lambda\mu_{bi}} + \frac{\nu_i \mu_{\underline{bo}} (S - \mu_{si})}{\underline{\kappa} + \overline{\kappa} + \lambda\mu_{bi}}. \tag{B.15}$$

The new system consists of Equations (5), (B.9), (B.13), and (B.15). These are six equations (because some are for each asset), and the six unknowns are $\mu_{\underline{bo}}$, $\mu_{\overline{b}}$, and $\{\mu_{bi}, \mu_{si}\}_{i\in\{1,2\}}$. Once this system is solved, the other measures can be computed as follows: $\mu_{\underline{s}i}$ from (B.10), $\mu_{\underline{n}i}$ from (B.11), $\mu_{\underline{b}i}$ from (B.12), $\mu_{\overline{s}i}$ from (B.14), $\mu_{\overline{\ell}i}$ from (3), and $\mu_{\overline{n}i}$ from (4).

To cover the case where search frictions are small, we make the change of variables $\varepsilon \equiv 1/\lambda$, $n \equiv \nu/\lambda$, $\alpha_i \equiv \nu_i/\nu$, $\gamma_{si} \equiv \lambda\mu_{si}$, and $\gamma_{\underline{bo}} \equiv \nu\mu_{\underline{bo}}$. Under the new variables, Equations (5), (B.9), (B.13), and (B.15) become

$$\overline{F} = \overline{\kappa}\mu_{\overline{b}} + \sum_{i=1}^{2} \mu_{\overline{b}}\gamma_{si}, \tag{B.16}$$

$$\underline{F} = \frac{\varepsilon\underline{\kappa}\gamma_{\underline{bo}}}{n} + \frac{\underline{\kappa}}{\overline{\kappa}+\underline{\kappa}} \sum_{i=1}^{2} \alpha_i\gamma_{\underline{bo}}(S - \varepsilon\gamma_{si}), \tag{B.17}$$

$$\mu_{bi} = \mu_{\overline{b}} + \frac{\underline{\kappa}\mu_{bi}\alpha_i\gamma_{\underline{bo}}(S-\varepsilon\gamma_{si})}{(\overline{\kappa}+\underline{\kappa})\left[\varepsilon(\overline{\kappa}+\underline{\kappa})+\mu_{bi}\right](\overline{\kappa}+\gamma_{si})}, \tag{B.18}$$

$$\gamma_{si} = \frac{\overline{\kappa}S}{\varepsilon\overline{\kappa}+\mu_{bi}} + \frac{\alpha_i\gamma_{\underline{bo}}(S-\varepsilon\gamma_{si})}{\varepsilon(\underline{\kappa}+\overline{\kappa})+\mu_{bi}}, \tag{B.19}$$

respectively.

B.1 Existence and Uniqueness

We next show that the system of Equations (B.16)-(B.19) has a unique symmetric solution when $\alpha_1 = \alpha_2 = 1$ (the "symmetric" case), and a unique solution when $\alpha_1 = 1$ and $\alpha_2 = 0$ (the "asymmetric" case). Using Equation (B.18) to eliminate $\gamma_{\underline{bo}}$ in Equation (B.19), we find

$$\gamma_{si} = \frac{\overline{\kappa}S}{\varepsilon\overline{\kappa}+\mu_{bi}} + (\mu_{bi} - \mu_{\overline{b}})\frac{(\overline{\kappa}+\underline{\kappa})(\overline{\kappa}+\gamma_{si})}{\underline{\kappa}\mu_{bi}}.$$

Multiplying by μ_{bi}, and setting $i = 1$, we find

$$\gamma_{s1}\mu_{\overline{b}} = \frac{\overline{\kappa}S\mu_{b1}}{\varepsilon\overline{\kappa}+\mu_{b1}} + (\mu_{b1} - \mu_{\overline{b}})\frac{\overline{\kappa}}{\underline{\kappa}}(\overline{\kappa}+\underline{\kappa}+\gamma_{s1}). \tag{B.20}$$

In the rest of the proof, we use Equations (B.16), (B.17), (B.18) for $i \in \{1,2\}$, and (B.19) for $i = 2$, to determine $\mu_{\overline{b}}$ and μ_{b1} as functions of $\gamma_{s1} \in (0, S/\varepsilon)$. We then plug these functions into Equation (B.20), and show that the resulting equation in the single unknown γ_{s1} has a unique solution.

We first solve for $\mu_{\overline{b}}$. In the asymmetric case, Equation (B.18) implies that $\mu_{b2} = \mu_{\overline{b}}$, Equation (B.19) implies that $\gamma_{s2} = \overline{\kappa}S/(\varepsilon\overline{\kappa} + \mu_{\overline{b}})$, and Equation (B.16) implies that

$$\overline{F} = \overline{\kappa}\mu_{\overline{b}} + \mu_{\overline{b}}\left(\gamma_{s1} + \frac{\overline{\kappa}S}{\varepsilon\overline{\kappa} + \mu_{\overline{b}}}\right). \tag{B.21}$$

The RHS of Equation (B.21) is (strictly) increasing in $\mu_{\overline{b}} \in (0, \infty)$, is equal to zero for $\mu_{\overline{b}} = 0$, and goes to ∞ for $\mu_{\overline{b}} \to \infty$. Therefore, Equation (B.21) has a unique solution $\mu_{\overline{b}} \in (0, \infty)$. This solution is decreasing in γ_{s1} because the RHS is increasing in γ_{s1}. In the symmetric case, Equation (B.16) implies that $\mu_{\overline{b}} = \overline{F}/(\overline{\kappa} + 2\gamma_{s1})$. This solution is again decreasing in γ_{s1}.

We next solve for μ_{b1}. Equation (B.17) implies that

$$\gamma_{\underline{bo}} = \frac{\underline{F}}{\frac{\varepsilon\underline{\kappa}}{n} + \frac{\underline{\kappa}}{\overline{\kappa}+\underline{\kappa}}\sum_{i=1}^{2}\alpha_i(S - \varepsilon\gamma_{si})} = \frac{\underline{F}}{\frac{\varepsilon\underline{\kappa}}{n} + \frac{\underline{\kappa}}{\overline{\kappa}+\underline{\kappa}}(1+\alpha_2)(S - \varepsilon\gamma_{s1})},$$

where the second step follows because in the symmetric case $\gamma_{s2} = \gamma_{s1}$ and in the asymmetric case $\alpha_2 = 0$. Plugging into Equation (B.18), setting $i = 1$, and dividing by μ_{b1}, we find

$$1 = \frac{\mu_{\overline{b}}}{\mu_{b1}} + \frac{(S - \varepsilon\gamma_{s1})n\underline{F}}{\left[\varepsilon(\overline{\kappa} + \underline{\kappa}) + \mu_{b1}\right](\overline{\kappa} + \gamma_{s1})\left[\varepsilon(\overline{\kappa} + \underline{\kappa}) + n(1+\alpha_2)(S - \varepsilon\gamma_{s1})\right]}. \tag{B.22}$$

The RHS of Equation (B.22) is decreasing in $\mu_{b1} \in (0, \infty)$, goes to ∞ for $\mu_{b1} \to 0$, and goes to zero for $\mu_{b1} \to \infty$. Therefore, Equation (B.21) has a unique solution $\mu_{b1} \in (0, \infty)$. This solution is decreasing in γ_{s1} because the RHS is decreasing in γ_{s1} and increasing in $\mu_{\overline{b}}$ (which is decreasing in γ_{s1}).

We next substitute $\mu_{\overline{b}}$ and μ_{b1} into Equation (B.20), and treat it as an equation in the single unknown γ_{s1}. To show uniqueness, we will show that the LHS is increasing in γ_{s1} and the RHS is decreasing. In the symmetric case, the LHS is equal to

$$\gamma_{s1}\mu_{\overline{b}} = \frac{\gamma_{s1}\overline{F}}{\overline{\kappa} + 2\gamma_{s1}},$$

and is increasing. In the asymmetric case, Equation (B.21) implies that the LHS is equal to

$$\gamma_{s1}\mu_{\overline{b}} = \overline{F} - \overline{\kappa}\mu_{\overline{b}} - \frac{\overline{\kappa}S\mu_{\overline{b}}}{\varepsilon\overline{\kappa} + \mu_{\overline{b}}},$$

and is increasing because $\mu_{\overline{b}}$ is decreasing in γ_{s1}. The first term in the RHS is increasing in μ_{b1}, and thus decreasing in γ_{s1}. To show that the second term is also decreasing, we multiply Equation (B.22) by $\mu_{b1}(\overline{\kappa}+\underline{\kappa}+\gamma_{s1})$:

$$(\mu_{b1}-\mu_{\overline{b}})(\overline{\kappa}+\underline{\kappa}+\gamma_{s1}) = \frac{\mu_{b1}(\overline{\kappa}+\underline{\kappa}+\gamma_{s1})(S-\varepsilon\gamma_{s1})n\underline{F}}{[\varepsilon(\overline{\kappa}+\underline{\kappa})+\mu_{b1}]\,(\overline{\kappa}+\gamma_{s1})\,[\varepsilon(\overline{\kappa}+\underline{\kappa})+n(1+\alpha_2)(S-\varepsilon\gamma_{s1})]}.$$

The RHS of this equation is decreasing in γ_{s1} because it is decreasing in γ_{s1} and increasing in μ_{b1} (which is decreasing in γ_{s1}). Therefore, the second term in the RHS of Equation (B.20) is decreasing in γ_{s1}.

To show existence, we note that for $\gamma_{s1}=0$, the LHS of Equation (B.20) is equal to zero, while the RHS is positive. Moreover, for $\gamma_{s1}=S/\varepsilon$, the LHS is equal to $S\mu_{\overline{b}}/\varepsilon$, while the RHS is equal to

$$\frac{\overline{\kappa}S\mu_{\overline{b}}}{\varepsilon\overline{\kappa}+\mu_{\overline{b}}} < \frac{S\mu_{\overline{b}}}{\varepsilon}$$

because $\mu_{b1}=\mu_{\overline{b}}$. Therefore, there exists a solution $\gamma_{s1}\in(0,S/\varepsilon)$.

B.2 Small Search Frictions

The case of small search frictions corresponds to small ε. Thus, the solution in this case is close to that for $\varepsilon=0$ provided that continuity holds. Our proof so far covers the case $\varepsilon=0$, except for existence. We next show that Assumption 2 ensures existence for $\varepsilon=0$. We also compute the solution in closed form and show continuity.

To emphasize that $\varepsilon=0$ is a limit case, we use m and g instead of μ and γ. Equations (B.16)-(B.19) become

$$\overline{F} = \overline{\kappa}m_{\overline{b}} + \sum_{i=1}^{2} m_{\overline{b}}g_{si}, \tag{B.23}$$

$$\underline{F} = \frac{\underline{\kappa}}{\overline{\kappa}+\underline{\kappa}}\sum_{i=1}^{2}\alpha_i g_{\underline{bo}}S, \tag{B.24}$$

$$m_{bi} = m_{\overline{b}} + \frac{\underline{\kappa}\alpha_i g_{\underline{bo}}S}{(\overline{\kappa}+\underline{\kappa})(\overline{\kappa}+g_{si})}, \tag{B.25}$$

$$g_{si} = \frac{\overline{\kappa}S}{m_{bi}} + \frac{\alpha_i g_{\underline{bo}}S}{m_{bi}}. \tag{B.26}$$

We first solve the system of (B.23)-(B.26) in the symmetric case ($\alpha_1 = \alpha_2 = 1$), suppressing the asset subscript because of symmetry. Equation (B.24) implies that

$$g_{\underline{bo}} = \frac{(\overline{\kappa} + \underline{\kappa})\underline{F}}{2\underline{\kappa}S}, \tag{B.27}$$

Equation (B.26) implies that

$$g_s = \frac{\overline{\kappa}S + \frac{\overline{\kappa}+\underline{\kappa}}{2\underline{\kappa}}\underline{F}}{m_b}, \tag{B.28}$$

and Equation (B.23) implies that

$$m_{\overline{b}} = \frac{\overline{F}}{\overline{\kappa} + 2g_s}. \tag{B.29}$$

Substituting $g_{\underline{bo}}$, g_s, and $m_{\overline{b}}$ from Equations (B.27)-(B.29) into Equation (B.25), we find that m_b solves the equation

$$1 = \frac{\overline{F}}{\overline{\kappa}m_b + 2\overline{\kappa}S + \frac{\overline{\kappa}+\underline{\kappa}}{\underline{\kappa}}\underline{F}} + \frac{\underline{F}}{2\overline{\kappa}m_b + 2\overline{\kappa}S + \frac{\overline{\kappa}+\underline{\kappa}}{\underline{\kappa}}\underline{F}}. \tag{B.30}$$

This equation has a positive solution because of Assumption 2.

We next consider the asymmetric case ($\alpha_1 = 1$, $\alpha_2 = 0$), and use $\hat{m}$ and $\hat{g}$ instead of m and g. Equation (B.25) implies that $\hat{m}_{b2} = \hat{m}_{\overline{b}}$, Equation (B.26) implies that

$$\hat{g}_{s2} = \frac{\overline{\kappa}S}{\hat{m}_{\overline{b}}}, \tag{B.31}$$

Equation (B.24) implies that

$$\hat{g}_{\underline{bo}} = \frac{(\overline{\kappa} + \underline{\kappa})\underline{F}}{\underline{\kappa}S}, \tag{B.32}$$

Equation (B.26) implies that

$$\hat{g}_{s1} = \frac{\overline{\kappa}S + \frac{\overline{\kappa}+\underline{\kappa}}{\underline{\kappa}}\underline{F}}{\hat{m}_{b1}}, \tag{B.33}$$

and Equation (B.23) implies that

$$\hat{m}_{\overline{b}} = \frac{\overline{F} - \overline{\kappa}S}{\overline{\kappa} + \hat{g}_{s1}}. \tag{B.34}$$

Substituting $\hat{g}_{\underline{b}o}$, $\hat{g}_{s1}$, and $\hat{m}_{\overline{b}}$ from Equations (B.32)-(B.34) into Equation (B.25), we find

$$\hat{m}_{b1} = \frac{\overline{F}}{\overline{\kappa}} - 2S - \frac{\underline{F}}{\underline{\kappa}}, \tag{B.35}$$

which is positive because of Assumption 2.

To show continuity at $\varepsilon = 0$, we write Equation (B.20) as

$$\gamma_{s1}\mu_{\overline{b}} - \frac{\overline{\kappa}S\mu_{b1}}{\varepsilon\overline{\kappa} + \mu_{b1}} - (\mu_{b1} - \mu_{\overline{b}})\frac{\overline{\kappa}}{\underline{\kappa}}(\overline{\kappa} + \underline{\kappa} + \gamma_{s1}) = 0,$$

and denote by $R(\gamma_{s1}, \varepsilon)$ the RHS (treating $\mu_{\overline{b}}$ and μ_{b1} as functions of $(\gamma_{s1}, \varepsilon)$). Because $\mu_{\overline{b}}, \mu_{b1} > 0$ for $(\gamma_{s1}, \varepsilon) = (g_{s1}, 0)$ (symmetric case) and $(\gamma_{s1}, \varepsilon) = (\hat{g}_{s1}, 0)$ (asymmetric case), the functions $\mu_{\overline{b}}$ and μ_{b1} are continuously differentiable around that point, and so is the function $R(\gamma_{s1}, \varepsilon)$. Moreover, our uniqueness proof shows that the derivative of $R(\gamma_{s1}, \varepsilon)$ w.r.t. γ_{s1} is positive. Therefore, the Implicit Function Theorem ensures that for small ε, Equation (B.20) has a continuous solution $\gamma_{s1}(\varepsilon)$. Because of uniqueness, this solution coincides with the one that we have identified.

C Utilities and Prices

We denote by C_i the cash collateral seized by the lender when the borrower cannot deliver instantly. The equations for the types' utilities are

$$\begin{aligned}
rV_{\overline{\ell}i} &= \delta + \overline{x} - y + \overline{\kappa}(V_{\overline{s}i} - V_{\overline{\ell}i}) + \nu_i\mu_{\underline{b}o}(V_{\overline{n}\underline{s}i} - V_{\overline{\ell}i}) && \text{(C.1)}\\
rV_{\overline{n}\underline{s}i} &= \delta + \overline{x} - y + w_i + \overline{\kappa}(V_{\overline{s}i} - V_{\overline{n}\underline{s}i}) + \underline{\kappa}(V_{\overline{\ell}i} - V_{\overline{n}\underline{s}i}) + \lambda\mu_{bi}(V_{\overline{n}\underline{n}i} - V_{\overline{n}\underline{s}i}) && \text{(C.2)}\\
rV_{\overline{n}\underline{n}i} &= \delta + \overline{x} - y + w_i + \overline{\kappa}(C_i - V_{\overline{n}\underline{n}i}) + \underline{\kappa}(V_{\overline{n}\underline{b}i} - V_{\overline{n}\underline{n}i}) && \text{(C.3)}\\
rV_{\overline{n}\underline{b}i} &= \delta + \overline{x} - y + w_i + \overline{\kappa}(C_i - V_{\overline{n}\underline{b}i}) + \lambda\mu_{si}(V_{\overline{\ell}i} - V_{\overline{n}\underline{b}i}) && \text{(C.4)}\\
rV_{\overline{s}i} &= \delta - y + \lambda\mu_{bi}(p_i - V_{\overline{s}i}) && \text{(C.5)}\\
rV_{\underline{b}o} &= -\underline{\kappa}V_{\underline{b}o} + \sum_{i=1}^{2}\nu_i\mu_{\overline{\ell}i}(V_{\underline{s}i} - V_{\underline{b}o}) && \text{(C.6)}\\
rV_{\underline{s}i} &= -w_i + \overline{\kappa}(V_{\underline{b}o} - V_{\underline{s}i}) - \underline{\kappa}V_{\underline{s}i} + \lambda\mu_{bi}(V_{\underline{n}i} + p_i - V_{\underline{s}i}) && \text{(C.7)}\\
rV_{\underline{n}i} &= -\delta + \underline{x} - y - w_i + \overline{\kappa}(V_{\underline{b}o} - C_i - V_{\underline{n}i}) + \underline{\kappa}(V_{\underline{b}i} - V_{\underline{n}i}) && \text{(C.8)}\\
rV_{\underline{b}i} &= -\delta - y - w_i + \overline{\kappa}(-C_i - V_{\underline{b}i}) + \lambda\mu_{si}(-p_i - V_{\underline{b}i}). && \text{(C.9)}
\end{aligned}$$

Using Equations (6)-(7) and (C.1)-(C.9), we will compute the lending fee w_i and the price p_i as a function of the short-selling surplus Σ_i. We will then derive a linear system for Σ_1 and Σ_2.

C.1 Lending Fee

Subtracting Equation (C.1) from (C.2), we find

$$(r+\overline{\kappa}+\underline{\kappa}+\nu_i\mu_{\underline{b}o})(V_{\overline{n}\underline{s}i}-V_{\overline{\ell}i})=w_i+\lambda\mu_{bi}(V_{\overline{n}\underline{n}i}-V_{\overline{n}\underline{s}i}), \tag{C.10}$$

subtracting (C.2) from (C.3), we find

$$(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})(V_{\overline{n}\underline{n}i}-V_{\overline{n}\underline{s}i})=\overline{\kappa}(C_i-V_{\overline{s}i})+\underline{\kappa}(V_{\overline{n}\underline{b}i}-V_{\overline{\ell}i}), \tag{C.11}$$

and subtracting (C.3) from (C.4), we find

$$(r+\overline{\kappa}+\underline{\kappa})(V_{\overline{n}\underline{b}i}-V_{\overline{n}\underline{n}i})=\lambda\mu_{si}(V_{\overline{\ell}i}-V_{\overline{n}\underline{b}i}). \tag{C.12}$$

Equations (C.11) and (C.12) imply that

$$V_{\overline{n}\underline{b}i}-V_{\overline{n}\underline{s}i}=\frac{\overline{\kappa}}{r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi}}(C_i-V_{\overline{s}i})+\frac{\underline{\kappa}(r+\overline{\kappa}+\underline{\kappa})-\lambda\mu_{si}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})}{(r+\overline{\kappa}+\underline{\kappa})(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})}(V_{\overline{n}\underline{b}i}-V_{\overline{\ell}i}).$$

Adding $V_{\overline{n}\underline{s}i}-V_{\overline{\ell}i}$ on both sides and solving for $V_{\overline{n}\underline{b}i}-V_{\overline{\ell}i}$, we find

$$V_{\overline{n}\underline{b}i}-V_{\overline{\ell}i}=\frac{(r+\overline{\kappa}+\underline{\kappa})(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})}{(r+\overline{\kappa}+\underline{\kappa})(r+\overline{\kappa}+\lambda\mu_{bi})+\lambda\mu_{si}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})}\left[\frac{\overline{\kappa}(C_i-V_{\overline{s}i})}{r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi}}+V_{\overline{n}\underline{s}i}-V_{\overline{\ell}i}\right].$$

Substituting $V_{\overline{n}\underline{b}i}-V_{\overline{\ell}i}$ from this equation into (C.11), we find

$$V_{\overline{n}\underline{n}i}-V_{\overline{n}\underline{s}i}=\frac{\overline{\kappa}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{si})(C_i-V_{\overline{s}i})+\underline{\kappa}(r+\overline{\kappa}+\underline{\kappa})(V_{\overline{n}\underline{s}i}-V_{\overline{\ell}i})}{(r+\overline{\kappa}+\underline{\kappa})(r+\overline{\kappa}+\lambda\mu_{bi})+\lambda\mu_{si}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})}.$$

Substituting $V_{\overline{n}\underline{n}i}-V_{\overline{n}\underline{s}i}$ from this equation into (C.10), and using (7), we can determine the lending fee as a function of the short-selling surplus:

$$\begin{aligned}&\left[r+\overline{\kappa}+\underline{\kappa}\frac{(r+\overline{\kappa})(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{si})+\lambda\mu_{si}(\underline{\kappa}+\lambda\mu_{bi})}{(r+\overline{\kappa}+\lambda\mu_{bi})(r+\overline{\kappa}+\underline{\kappa})+\lambda\mu_{si}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})}+\nu_i\mu_{\underline{b}o}\right]\theta\Sigma_i\\ =\;& w_i+\frac{\overline{\kappa}\lambda\mu_{bi}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{si})}{(r+\overline{\kappa}+\underline{\kappa})(r+\overline{\kappa}+\lambda\mu_{bi})+\lambda\mu_{si}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})}(C_i-V_{\overline{s}i}).\end{aligned} \tag{C.13}$$

C.2 Price

Equation (C.5) implies that

$$V_{\overline{s}i} - p_i = \frac{\delta - y - rp_i}{r + \lambda\mu_{bi}}. \tag{C.14}$$

Subtracting rp_i from both sides of (C.1), and using (7) and (C.14), we find

$$V_{\overline{\ell}i} - p_i = \frac{1}{r + \overline{\kappa}} \left[\delta + \overline{x} - y - rp_i + \nu_i \mu_{\underline{bo}} \theta \Sigma_i + \overline{\kappa} \frac{\delta - y - rp_i}{r + \lambda\mu_{bi}} \right]. \tag{C.15}$$

Substituting (C.14) and (C.15) into (9), we find

$$\delta - y - rp_i + \frac{(1-\phi)(r + \lambda\mu_{bi})}{r + \overline{\kappa} + (1-\phi)\lambda\mu_{bi}} \left[\overline{x} + \nu_i \mu_{\underline{bo}} \theta \Sigma_i - (r + \overline{\kappa}) V_{\overline{b}} \right] = 0. \tag{C.16}$$

Substituting $d - y - rp_i$ from (C.16) into (C.15), we find

$$V_{\overline{\ell}i} - p_i = \frac{\phi(\overline{x} + \nu_i \mu_{\underline{bo}} \theta \Sigma_i) + (1-\phi)(r + \overline{\kappa} + \lambda\mu_{bi}) V_{\overline{b}}}{r + \overline{\kappa} + (1-\phi)\lambda\mu_{bi}}.$$

Substituting $V_{\overline{\ell}i} - p_i$ from this equation into (6) and solving for $V_{\overline{b}}$, we find

$$V_{\overline{b}} = \frac{\phi \sum_{j=1}^{2} \frac{\lambda\mu_{sj}}{r + \overline{\kappa} + (1-\phi)\lambda\mu_{bj}} (\overline{x} + \nu_j \mu_{\underline{bo}} \theta \Sigma_j)}{(r + \overline{\kappa}) \left[1 + \phi \sum_{j=1}^{2} \frac{\lambda\mu_{sj}}{r + \overline{\kappa} + (1-\phi)\lambda\mu_{bj}} \right]}.$$

Substituting $V_{\overline{b}}$ from this equation into (C.16), we can determine the price as a function of the short-selling surplus:

$$p_i = \frac{\delta - y}{r} + \frac{(1-\phi)(r + \lambda\mu_{bi})}{r \left[r + \overline{\kappa} + (1-\phi)\lambda\mu_{bi} \right]} \left[\overline{x} + \nu_i \mu_{\underline{bo}} \theta \Sigma_i - \frac{\phi \sum_{j=1}^{2} \frac{\lambda\mu_{sj}}{r + \overline{\kappa} + (1-\phi)\lambda\mu_{bj}} (\overline{x} + \nu_j \mu_{\underline{bo}} \theta \Sigma_j)}{1 + \phi \sum_{j=1}^{2} \frac{\lambda\mu_{sj}}{r + \overline{\kappa} + (1-\phi)\lambda\mu_{bj}}} \right]. \tag{C.17}$$

C.3 Short-Selling Surplus

Adding Equations (C.2) and (C.7), and subtracting Equations (C.6) and (C.1), we find

$$(r + \overline{\kappa} + \underline{\kappa} + \nu_i \mu_{\underline{bo}} \theta) \Sigma_i + \sum_{j=1}^{2} \nu_j \mu_{\overline{\ell}j} (1-\theta) \Sigma_j = \lambda\mu_{bi} (V_{\overline{n}\underline{n}i} + V_{\underline{n}i} + p_i - V_{\overline{n}\underline{s}i} - V_{\underline{s}i}). \tag{C.18}$$

Adding Equations (C.3), (C.8), and $rp_i = rp_i$, and subtracting Equations (C.2) and (C.7), we find

$$(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})(V_{\overline{n}\underline{n}i}+V_{\underline{n}i}+p_i-V_{\overline{n}\underline{s}i}-V_{\underline{s}i}) = rp_i-\delta+\underline{x}-y+\overline{\kappa}(p_i-V_{\overline{s}i})+\underline{\kappa}(V_{\overline{n}\underline{b}i}+V_{\underline{b}i}+p_i-V_{\overline{\ell}i}). \quad \text{(C.19)}$$

Adding Equations (C.4), (C.9), and $rp_i = rp_i$, and subtracting Equation (C.1), we find

$$(r+\overline{\kappa}+\lambda\mu_{si})(V_{\overline{n}\underline{b}i}+V_{\underline{b}i}+p_i-V_{\overline{\ell}i}) = rp_i-\delta-y+\overline{\kappa}(p_i-V_{\overline{s}i})-\nu_i\mu_{\underline{bo}}\theta\Sigma_i. \quad \text{(C.20)}$$

Substituting $V_{\overline{n}\underline{b}i}+V_{\underline{b}i}+p_i-V_{\overline{\ell}i}$ from Equation (C.20) into (C.19), and then substituting $V_{\overline{n}\underline{n}i}+V_{\underline{n}i}+p_i-V_{\overline{n}\underline{s}i}-V_{\underline{s}i}$ from Equation (C.19) into (C.18), we find

$$\begin{aligned}
&\left[r+\overline{\kappa}+\underline{\kappa}+\nu_i\mu_{\underline{bo}}\theta\left[1+\frac{\lambda\mu_{bi}\underline{\kappa}}{(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})(r+\overline{\kappa}+\lambda\mu_{si})}\right]\right]\Sigma_i+\sum_{j=1}^{2}\nu_j\mu_{\overline{\ell}j}(1-\theta)\Sigma_j \\
= \quad &\frac{\lambda\mu_{bi}}{r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi}}\left[\underline{x}+\frac{r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{si}}{r+\overline{\kappa}+\lambda\mu_{si}}\left[rp_i-\delta-y+\overline{\kappa}(p_i-V_{\overline{s}i})\right]\right]. \quad \text{(C.21)}
\end{aligned}$$

To derive an equation involving only Σ_1 and Σ_2, we need to eliminate the price p_i. We have

$$\begin{aligned}
&rp_i-\delta-y+\overline{\kappa}(p_i-V_{\overline{s}i}) \\
= \quad &-2y+rp_i-\delta+y+\overline{\kappa}\frac{rp_i-\delta+y}{r+\lambda\mu_{bi}} \\
= \quad &-2y+\frac{(1-\phi)(r+\overline{\kappa}+\lambda\mu_{bi})}{r+\overline{\kappa}+(1-\phi)\lambda\mu_{bi}}\left[\overline{x}+\nu_i\mu_{\underline{bo}}\theta\Sigma_i-\frac{\phi\sum_{j=1}^{2}\frac{\lambda\mu_{sj}}{r+\overline{\kappa}+(1-\phi)\lambda\mu_{bj}}(\overline{x}+\nu_j\mu_{\underline{bo}}\theta\Sigma_j)}{1+\phi\sum_{j=1}^{2}\frac{\lambda\mu_{sj}}{r+\overline{\kappa}+(1-\phi)\lambda\mu_{bj}}}\right],
\end{aligned}$$

where the first step follows from Equation (C.14) and the second from (C.17). Plugging back into Equation (C.21), we can write it as

$$a_i\Sigma_i+\sum_{j=1}^{2}f_j\Sigma_j+b_i\sum_{j=1}^{2}g_j\Sigma_j = c_i, \quad \text{(C.22)}$$

where

$$a_i = r+\overline{\kappa}+\underline{\kappa}+\nu_i\mu_{\underline{bo}}\theta\left[\frac{r+\overline{\kappa}+\underline{\kappa}}{r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi}}+\frac{\phi(r+\overline{\kappa})\lambda\mu_{bi}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{si})}{(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})(r+\overline{\kappa}+\lambda\mu_{si})[r+\overline{\kappa}+(1-\phi)\lambda\mu_{bi}]}\right],$$

$$f_i = \nu_i\mu_{\overline{\ell}i}(1-\theta),$$

$$b_i = \frac{(1-\phi)\lambda\mu_{bi}(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{si})(r+\overline{\kappa}+\lambda\mu_{bi})}{(r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi})(r+\overline{\kappa}+\lambda\mu_{si})\left[r+\overline{\kappa}+\lambda(1-\phi)\lambda\mu_{bi}\right]},$$

$$g_i = \phi\nu_i\mu_{\underline{bo}}\theta\frac{\frac{\lambda\mu_{si}}{r+\overline{\kappa}+(1-\phi)\lambda\mu_{bi}}}{1+\phi\sum_{j=1}^{2}\frac{\lambda\mu_{sj}}{r+\overline{\kappa}+(1-\phi)\lambda\mu_{bj}}},$$

$$c_i = \frac{\lambda\mu_{bi}}{r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{bi}}\left[\underline{x}-\frac{r+\overline{\kappa}+\underline{\kappa}+\lambda\mu_{si}}{r+\overline{\kappa}+\lambda\mu_{si}}\left[2y-\frac{(1-\phi)(r+\overline{\kappa}+\lambda\mu_{bi})}{r+\overline{\kappa}+(1-\phi)\lambda\mu_{bi}}\frac{\overline{x}}{1+\phi\sum_{j=1}^{2}\frac{\lambda\mu_{sj}}{r+\overline{\kappa}+(1-\phi)\lambda\mu_{sj}}}\right]\right].$$

The short-selling surpluses Σ_1 and Σ_2 are the solution to the linear system consisting of Equation (C.22) for $i \in \{1,2\}$.

Note that the collateral C_i does not enter in Equation (C.22), and thus does not affect the short-selling surplus. It neither affects the price, from Equation (C.17). It affects only the lending fee because when lenders can seize more collateral they accept a lower fee. From now on (and as stated in Footnote 13), we set the collateral equal to the utility of a seller $\overline{s}i$, i.e.,

$$C_i = V_{\overline{s}i}. \tag{C.23}$$

D Proofs of Propositions 5-11

Proof of Proposition 5: From Appendix B we know that given the short-selling decisions $\nu_1 = \nu_2 = \nu$, the population measures are uniquely determined. From Appendix C we know that given any short-selling decisions and population measures, the utilities, prices, and lending fees are uniquely determined. Therefore, what is left to show is (i) the short-selling surplus Σ is positive, (ii) buyers' and sellers' reservation values are ordered as in Equation (8), and (iii) agents' trading strategies are optimal. To show these results, we recall from Appendix B that when search frictions become small, i.e., λ goes to ∞ holding $n \equiv \nu/\lambda$ constant, μ_b converges to m_b, $\mu_{\overline{\ell}}$ converges to S, $\lambda\mu_s$ converges to g_s, and $\nu\mu_{\underline{bo}}$ converges to $g_{\underline{bo}}$.

We start by computing Σ, w, and p, thus proving Proposition 6. Equation (C.22) implies that when $\Sigma_1 = \Sigma_2 \equiv \Sigma$,

$$\Sigma = \frac{c}{a+2(f+bg)},$$

where we suppress the asset subscripts from a, b, c, f, g because of symmetry. When search frictions

become small, a and b converge to positive limits, c converges to

$$\underline{x} - \frac{r + \overline{\kappa} + \underline{\kappa} + g_s}{r + \overline{\kappa} + g_s}(2y - \overline{x}), \tag{D.1}$$

g converges to zero, and f converges to ∞, being asymptotically equal to $\nu S(1-\theta)$. Therefore, the surplus converges to zero, and its asymptotic behavior is as in Proposition 6.

Equations (C.13) and (C.23) imply that the lending fee is

$$w = \left[r + \overline{\kappa} + \underline{\kappa} \frac{(r + \overline{\kappa})(r + \overline{\kappa} + \underline{\kappa} + \lambda\mu_s) + \lambda\mu_s(\underline{\kappa} + \lambda\mu_b)}{(r + \overline{\kappa} + \lambda\mu_b)(r + \overline{\kappa} + \underline{\kappa}) + \lambda\mu_s(r + \overline{\kappa} + \underline{\kappa} + \lambda\mu_b)} + \nu\mu_{\underline{bo}} \right] \theta\Sigma.$$

Because the term in brackets converges to

$$r + \overline{\kappa} + \underline{\kappa} \frac{g_s}{r + \overline{\kappa} + \underline{\kappa} + g_s} + g_{\underline{bo}},$$

the lending fee converges to zero, and its asymptotic behavior is as in Proposition 6.

Equation (C.17) implies that the price is equal to

$$p = \frac{\delta - y}{r} + \frac{1}{r}\left[1 - \frac{\phi r + \overline{\kappa}}{(1-\phi)\lambda m_b} + o(1/\lambda)\right]\left[\overline{x} + g_{\underline{bo}}\theta\Sigma - \frac{2\phi g_s \overline{x}}{(1-\phi)\lambda m_b} + o(1/\lambda)\right].$$

Using this equation and the fact that Σ is in order $1/\lambda$, it is easy to check that the asymptotic behavior (i.e., order $1/\lambda$) of the price is as in Proposition 6.

To show that Σ is positive, we need to show that (D.1) is positive. This follows because (11) implies that

$$\underline{x} > 2y + \frac{\underline{\kappa}}{r + \overline{\kappa} + g_s}(2y - \overline{x}) > 2y - \overline{x} + \frac{\underline{\kappa}}{r + \overline{\kappa} + g_s}(2y - \overline{x}) = \frac{r + \overline{\kappa} + \underline{\kappa} + g_s}{r + \overline{\kappa} + g_s}(2y - \overline{x}). \tag{D.2}$$

We next show that reservation values are ordered as in Equation (8), i.e., $\Delta_{\underline{b}} > \Delta_{\overline{b}}$ and $\Delta_{\overline{s}} > \Delta_{\underline{s}}$. For this, we need to compute $V_{\underline{b}}$ and $V_{\underline{n}} - V_{\underline{s}}$. Adding (C.9) and $rp = rp$, and using (C.23), we find

$$V_{\underline{b}} + p = \frac{rp - \delta - y - w + \overline{\kappa}(p - V_{\overline{s}})}{r + \overline{\kappa} + \lambda\mu_s}. \tag{D.3}$$

Adding (C.8) and $rp = rp$, and subtracting (C.7), we similarly find

$$V_{\underline{n}} + p - V_{\underline{s}} = \frac{rp - \delta + \underline{x} - y + \underline{\kappa}(V_{\underline{b}} + p) + \overline{\kappa}(p - V_{\overline{s}})}{r + \overline{\kappa} + \underline{\kappa} + \lambda\mu_b}. \tag{D.4}$$

Inequality $\Delta_{\underline{b}} > \Delta_{\overline{b}}$ is equivalent to

$$\begin{aligned} & -V_{\underline{b}} - p > V_{\overline{\ell}} - p - V_{\overline{b}} \\ \Leftrightarrow \quad & \frac{\delta + y - rp + w - \overline{\kappa}(p - V_{\overline{s}})}{r + \overline{\kappa} + \lambda\mu_s} > \frac{\phi}{1-\phi}(p - V_{\overline{s}}) \\ \Leftrightarrow \quad & \frac{\delta + y - rp + w - \overline{\kappa}\frac{rp - \delta + y}{r + \lambda\mu_b}}{r + \overline{\kappa} + \lambda\mu_s} > \frac{\phi}{1-\phi}\frac{rp - \delta + y}{r + \lambda\mu_b} \end{aligned} \tag{D.5}$$

where the second step follows from (9) and (D.3), and the third from (C.14). Because rp converges to $\delta + \overline{x} - y$, and w converges to zero, the LHS of (D.5) converges to $(2y - \overline{x})/(r + \overline{\kappa} + g_s)$, which is positive from Assumption 1, while the RHS converges to zero. Inequality $\Delta_{\overline{s}} > \Delta_{\underline{s}}$ is equivalent to

$$\begin{aligned} & V_{\underline{n}} + p - V_{\underline{s}} > p - V_{\overline{s}} \\ \Leftrightarrow \quad & \frac{\underline{x} + \frac{r + \overline{\kappa} + \underline{\kappa} + \lambda\mu_s}{r + \underline{\kappa} + \lambda\mu_s}[rp - \delta - y + \overline{\kappa}(p - V_{\overline{s}})] - \frac{\underline{\kappa}}{r + \overline{\kappa} + \lambda\mu_s} w}{r + \overline{\kappa} + \underline{\kappa} + \lambda\mu_b} > \frac{rp - \delta + y}{r + \lambda\mu_b}, \end{aligned}$$

where the second step follows from (C.14), (D.3), and (D.4). When search frictions become small, this inequality holds if the limit of the numerator in the LHS exceeds that for the RHS, i.e.,

$$\underline{x} - \frac{r + \overline{\kappa} + \underline{\kappa} + g_s}{r + \underline{\kappa} + g_s}(2y - \overline{x}) > \overline{x}.$$

This holds because of the first inequality in (D.2).

We finally show that trading strategies are optimal. The flow benefit that an average-valuation agent can derive from a long position in asset i is bounded above by $\delta - y + w$, and the flow benefit for a short position is bounded above by $-\delta - y$. Therefore, an average-valuation agent finds it optimal to establish no position, or to unwind a previously established one, if $(\delta - y + w)/r < \min\{p, C\}$ and $(\delta + y)/r > p$. These conditions are satisfied for small frictions because p converges to $(\delta + \overline{x} - y)/r$, w converges to zero, $C - p$ converges to zero, and $2y > \overline{x}$.

A high-valuation agent finds it optimal to buy asset i if $V_{\overline{\ell}} - p - V_{\overline{b}} \geq 0$. This condition is satisfied because

$$V_{\overline{\ell}} - p - V_{\overline{b}} = \frac{\phi}{1-\phi}(p - V_{\overline{s}}) = \frac{\phi}{1-\phi}\frac{rp - \delta + y}{r + \lambda\mu_b} \sim \frac{\phi}{1-\phi}\frac{\overline{x}}{\lambda\mu_b} \geq 0.$$

The agent finds it optimal to lend the asset because $V_{\overline{n}\underline{s}} - V_{\overline{\ell}} = \theta\Sigma > 0$. Likewise, a low-valuation agent finds it optimal to borrow asset i because $V_{\underline{s}} - V_{\underline{bo}} = (1-\theta)\Sigma > 0$, and to sell it because $V_{\underline{n}} + p - V_{\underline{s}} = p - \Delta_{\underline{s}} > p - \Delta_{\overline{s}} = p - V_{\overline{s}} > 0$. ∎

Proof of Proposition 6: See the proof of Proposition 5. ∎

Proof of Proposition 7: We need to show that (i) the short-selling surplus Σ_1 is positive and Σ_2 is negative, (ii) buyers' and sellers' reservation values are ordered as in Equation (8), and (iii) agents' trading strategies are optimal. We recall from Appendix B that for small search frictions and given the short-selling decisions $\nu_1 = \nu$ and $\nu_2 = 0$, μ_{bi} converges to $\hat{m}_{bi}$, $\mu_{\overline{\ell}i}$ converges to S, $\lambda\mu_{si}$ converges to $\hat{g}_{si}$, and $\nu\mu_{\underline{bo}}$ converges to $\hat{g}_{\underline{bo}}$.

We start by computing Σ_1, w_1, p_1, and p_2, thus proving Proposition 8. Equation (C.22) implies that when $\nu_2 = 0$,

$$\Sigma_1 = \frac{c_1}{a_1 + f_1 + b_1 g_1}.$$

When search frictions become small, c_1 converges to

$$\underline{x} - \frac{r + \overline{\kappa} + \underline{\kappa} + \hat{g}_{s1}}{r + \overline{\kappa} + \hat{g}_{s1}}(2y - \overline{x}), \tag{D.6}$$

and the dominant term in the denominator is $f_1 \sim \nu S(1-\theta)$. Therefore, the surplus converges to zero, and its asymptotic behavior is as in Proposition 8. To determine the asymptotic behavior of the lending fee and the price, we proceed as in the proof of Proposition 5.

To show that Σ_1 is positive, we need to show that (D.6) is positive. This follows from (D.2) and the fact that $\hat{g}_{s1} > g_s$, established in the proof of Proposition 9. To show that Σ_2 is negative, we note that from (C.22),

$$\Sigma_2 = \frac{c_2 - (f_1 + b_2 g_1)\Sigma_1}{a_2} = \frac{c_2 - \frac{f_1 + b_2 g_1}{a_1 + f_1 + b_1 g_1} c_1}{a_2}.$$

When search frictions become small, the numerator converges to the same limit as $c_2 - c_1$. This limit is equal to

$$\left[\frac{r + \overline{\kappa} + \underline{\kappa} + \hat{g}_{s1}}{r + \overline{\kappa} + \hat{g}_{s1}} - \frac{r + \overline{\kappa} + \underline{\kappa} + \hat{g}_{s2}}{r + \overline{\kappa} + \hat{g}_{s2}}\right](2y - \overline{x}),$$

and is negative if $\hat{g}_{s1} > \hat{g}_{s2}$. Using (B.31) and (B.33), we can write this inequality as

$$\frac{\overline{\kappa}S + \frac{\overline{\kappa}+\kappa}{\underline{\kappa}}\underline{F}}{\hat{m}_{b1}} > \frac{\overline{\kappa}S}{\hat{m}_{\overline{b}}}. \tag{D.7}$$

Equations (B.33)-(B.35) imply that

$$\hat{m}_{\overline{b}} = \frac{\overline{F} - \overline{\kappa}S}{\overline{F} - \overline{\kappa}S + \underline{F}}\hat{m}_{b1}. \tag{D.8}$$

Using this equation, we can write (D.7) as

$$\frac{\overline{\kappa}S + \frac{\overline{\kappa}+\kappa}{\underline{\kappa}}\underline{F}}{\overline{\kappa}S} > \frac{\overline{F} - \overline{\kappa}S + \underline{F}}{\overline{F} - \overline{\kappa}S}.$$

It is easy to check that this inequality holds because of Assumption 2.

To show that $\Delta_{\underline{b}i} > \Delta_{\overline{b}}$ and $\Delta_{\overline{s}i} > \Delta_{\underline{s}i}$, we proceed as in the proof of Proposition 5. The only change is that the condition for $\Delta_{\overline{s}i} > \Delta_{\underline{s}i}$ now is

$$\underline{x} - \frac{r + \overline{\kappa} + \underline{\kappa} + \hat{g}_s}{r + \underline{\kappa} + \hat{g}_s}(2y - \overline{x}) > \overline{x}.$$

This inequality is implied by the first inequality in (D.2) and the fact that $\hat{g}_{s1} > g_s$. Finally, the arguments in the proof of Proposition 5 establish that trading strategies are optimal. ∎

Proof of Proposition 8: See the proof of Proposition 7. ∎

Proof of Proposition 9: We start with a lemma.

Lemma 1. *For $\chi < 1$, inequality $(1-\chi)\hat{m}_{b1} > m_b$ is equivalent to*

$$(1-2\chi)(\underline{F} - \chi\overline{\kappa}\hat{m}_{b1}) > \chi\overline{F}. \tag{D.9}$$

Proof: Since m_b is the unique positive solution of (B.30), whose RHS is decreasing in m_b, inequality $(1-\chi)\hat{m}_{b1} > m_b$ is equivalent to

$$1 > \frac{\overline{F}}{\overline{\kappa}(1-\chi)\hat{m}_{b1} + 2\overline{\kappa}S + \frac{\overline{\kappa}+\kappa}{\underline{\kappa}}\underline{F}} + \frac{\underline{F}}{2\overline{\kappa}(1-\chi)\hat{m}_{b1} + 2\overline{\kappa}S + \frac{\overline{\kappa}+\kappa}{\underline{\kappa}}\underline{F}}$$

$$\Leftrightarrow \quad 1 > \frac{\overline{F}}{\overline{F} + \underline{F} - \chi\overline{\kappa}\hat{m}_{b1}} + \frac{\underline{F}}{\overline{F} + \underline{F} + (1-2\chi)\overline{\kappa}\hat{m}_{b1}}$$

$$\Leftrightarrow \quad \frac{\underline{F} - \chi\overline{\kappa}\hat{m}_{b1}}{\overline{F} + \underline{F} - \chi\overline{\kappa}\hat{m}_{b1}} > \frac{\underline{F}}{\overline{F} + \underline{F} + (1-2\chi)\overline{\kappa}\hat{m}_{b1}},$$

where the second step follows from (B.35). The last inequality implies (D.9). ∎

Result (i): We need to show that $\hat{m}_{b1} > m_b$ and $\hat{g}_{s1} > g_s$. Since (D.9) holds for $\chi = 0$, Lemma 1 implies that $\hat{m}_{b1} > m_b$. Using (B.28) and (B.33), we can write inequality $\hat{g}_{s1} > g_s$ as

$$\frac{\overline{\kappa}S + \frac{\overline{\kappa}+\underline{\kappa}}{2\underline{\kappa}}\underline{F}}{\overline{\kappa}S + \frac{\overline{\kappa}+\underline{\kappa}}{\underline{\kappa}}\underline{F}}\hat{m}_{b1} < m_b.$$

Using Lemma 1, we then need to show that

$$(1-2\chi)(\underline{F} - \chi\overline{\kappa}\hat{m}_{b1}) < \chi\overline{F}, \tag{D.10}$$

for

$$\chi = \frac{\frac{\overline{\kappa}+\underline{\kappa}}{2\underline{\kappa}}\underline{F}}{\overline{\kappa}S + \frac{\overline{\kappa}+\underline{\kappa}}{\underline{\kappa}}\underline{F}}.$$

Plugging for χ, we can write (D.10) as

$$\overline{\kappa}S(\underline{F} - \chi\overline{\kappa}\hat{m}_{b1}) < \frac{\overline{\kappa}+\underline{\kappa}}{2\underline{\kappa}}\underline{F}\overline{F},$$

which holds because of Assumption 2 and $\hat{m}_{b1} > 0$.

Result (ii): We need to show that $\hat{m}_{b2} < m_b$ and $\hat{g}_{s2} < g_s$. Using (D.8) and $\hat{m}_{b2} = \hat{m}_{\overline{b}}$, we can write inequality $\hat{m}_{b2} < m_b$ as

$$\frac{\overline{F} - \overline{\kappa}S}{\overline{F} - \overline{\kappa}S + \underline{F}}\hat{m}_{b1} < m_b.$$

Using Lemma 1, we then need to show Equation (D.10) for

$$\chi = \frac{\underline{F}}{\overline{F} - \overline{\kappa}S + \underline{F}}.$$

Plugging for χ, we can write (D.10) as

$$\frac{\overline{F} - \overline{\kappa}S - \underline{F}}{\overline{F} - \overline{\kappa}S + \underline{F}}\left(\overline{F} - \overline{\kappa}S + \underline{F} - \overline{\kappa}\hat{m}_{b1}\right) < \overline{F},$$

which holds because $\hat{m}_{b1} > 0$. Using (B.28), (B.31), and (D.8), we can write inequality $\hat{g}_{s2} < g_s$ as

$$\frac{\overline{F} - \overline{\kappa}S}{\overline{F} - \overline{\kappa}S + \underline{F}}\frac{\overline{\kappa}S + \frac{\overline{\kappa}+\underline{\kappa}}{2\underline{\kappa}}\underline{F}}{\overline{\kappa}S}\hat{m}_{b1} > m_b.$$

Using Lemma 1, we then need to show (D.9) for

$$\chi = \frac{\underline{F}}{\overline{\overline{F}} - \overline{\kappa}S + \underline{F}} \left(1 - \frac{\overline{\kappa} + \underline{\kappa}}{2\underline{\kappa}} \frac{\overline{F} - \overline{\kappa}S}{\overline{\kappa}S}\right).$$

Assumption 2 implies that

$$\chi < \frac{\underline{F}}{\overline{\overline{F}} - \overline{\kappa}S + \underline{F}} \left(1 - \frac{\overline{\kappa} + \underline{\kappa}}{2\underline{\kappa}}\right) < \frac{\underline{F}}{2(\overline{F} - \overline{\kappa}S + \underline{F})} \equiv \hat{\chi}.$$

Because $\hat{\chi}, \hat{m}_{b1} > 0$, Equation (D.9) holds for χ if it holds for $\hat{\chi}$. The latter is easy to check using Assumption 2.

Result (iii): Equations (14), (18), and $\hat{g}_{s1} > g_s$, imply that Σ_i in the symmetric equilibrium is smaller than Σ_1 in the asymmetric equilibrium. Since, in addition, $\hat{g}_{\underline{bo}} > g_{\underline{bo}}$ (from (B.27) and (B.32)), (13) and (17) imply that the lending fee w_i in the symmetric equilibrium is smaller than w_1 in the asymmetric equilibrium.

Result (iv): For $\phi = 0$, the result follows from (12), (15), $\hat{m}_{b1} > m_b > \hat{m}_{b2}$, $\hat{g}_{\underline{bo}} > g_{\underline{bo}}$, and the fact that Σ_i in the symmetric equilibrium is smaller than Σ_1 in the asymmetric equilibrium. An example where the prices of both assets are higher in the asymmetric equilibrium is $S = 0.5$, $\overline{F} = 3$, $\underline{F} = 5.7$, $\overline{\kappa} = 1$, $\underline{\kappa} = 3$, $\phi = \theta = 0.5$, $r = 4\%$, $\delta = 1$, $\overline{x} = 0.4$, $\underline{x} = 1.6$, $y = 0.5$, and any ν/λ. ∎

Proof of Proposition 10: We show that buying asset 2 and shorting asset 1 is unprofitable under

$$p_1 - p_2 < \frac{w_1}{r} + \frac{\overline{x}}{\lambda \hat{m}_{b1}} + \frac{\overline{\kappa x}}{r(\nu S + \lambda \hat{m}_{b2})}. \tag{D.11}$$

(which is implied by (19)), while buying asset 1 and shorting asset 2 is unprofitable under (20). We then show that Equations (19) and (20) are satisfied if ν/λ is in an interval (n_1, n_2).

Buy asset 2, short asset 1

Because trading opportunities arrive one at a time, an arbitrageur cannot set up the two legs of the position simultaneously. The arbitrageur can, for example, buy asset 2 first, then borrow asset 1, and then sell asset 1. Alternatively, he can borrow asset 1 first, then buy asset 2, and then sell asset 1. The final possibility, which is to sell asset 1 before buying asset 2 is suboptimal. Indeed, for small search frictions the time to meet a buyer converges to zero while the time to meet a seller does not. Therefore, the cost of being unhedged converges to zero only when asset 2 is bought before asset 1 is sold.

Suppose now that the arbitrage strategy is profitable. Because the payoff of the strategy is decreasing in asset 1's lending fee, there exists a fee $\overline{w}_1 > w_1$ for which the arbitrageur is indifferent between following the strategy and holding no position. If for this fee it is optimal to initiate the strategy by buying asset 2, the arbitrageur can be in three possible states:

- Long position in asset 2. State $n2$ with utility V_{n2}.
- Long position in asset 2 and borrowed asset 1. State $s1n2$ with utility V_{s1n2}.
- Long position in asset 2 and short in asset 1. State $n1n2$ with utility V_{n1n2}.

The utilities are characterized by the following flow-value equations:

$$rV_{n2} = \delta - y + \nu\mu_{\bar{\ell}1}(V_{s1n2} - V_{n2}) \tag{D.12}$$

$$rV_{s1n2} = \delta - y - \overline{w}_1 + \lambda\mu_{b1}(V_{n1n2} + p_1 - V_{s1n2}) + \overline{\kappa}(V_{n2} - V_{s1n2}) \tag{D.13}$$

$$rV_{n1n2} = -\overline{w}_1 + \overline{\kappa}(V_{n2} - C_1 - V_{n1n2}). \tag{D.14}$$

Solving (D.12)-(D.14), we find

$$rV_{n2} = \delta - y + \frac{\nu\mu_{\bar{\ell}1}}{r + \overline{\kappa} + \nu\mu_{\bar{\ell}1}}\left[-\overline{w}_1 + \frac{\lambda\mu_{b1}}{r + \overline{\kappa} + \lambda\mu_{b1}}\left[rp_1 - \delta + y + \overline{\kappa}(p_1 - C_1)\right]\right].$$

The arbitrageur is indifferent between initiating the strategy and holding no position if V_{n2} is equal to p_2. Using this condition, and substituting C_1 from Equations (C.14) and (C.23), we find

$$\overline{w}_1 = \frac{\lambda\mu_{b2}}{r + \lambda\mu_{b2}}(rp_1 - \delta + y) - \frac{r + \overline{\kappa} + \nu\mu_{\bar{\ell}2}}{\nu\mu_{\bar{\ell}2}}(rp_2 - \delta - y).$$

For small search frictions, this equation becomes

$$\overline{w}_1 = r(p_1 - p_2) - \frac{r\overline{x}}{\lambda\hat{m}_{b1}} - \frac{(r + \overline{\kappa})\overline{x}}{\nu S},$$

and is inconsistent with (D.11) since $w_1 < \overline{w}_1$.

Suppose instead that it is optimal to initiate the strategy by borrowing asset 1. The arbitrageur then starts from a state $s1$, in which he has borrowed asset 1 but holds no position in asset 2. The utility V_{s1} in this state is characterized by

$$rV_{s1} = -w_1 + \lambda\mu_{s2}(V_{s1n2} - p_2 - V_{s1}). \tag{D.15}$$

The utility in states $s1n2$ and $n1n2$ is given by (D.13) and (D.14), respectively. The utility in state $n2$, however, is given by

$$rV_{n2} = \delta - y + \nu\mu_{\overline{\ell}1}(V_{s1n2} - V_{n2}) + \lambda\mu_{b2}(p_2 - V_{n2}) \tag{D.16}$$

instead of (D.12). Indeed, since it suboptimal to initiate the strategy by buying asset 2, buying that asset is dominated by holding no position. Therefore, if the arbitrageur finds himself with a long position in asset 2, he prefers to unwind it upon meeting a seller. Equations (D.13), (D.14), and (D.16) imply that

$$V_{s1n2} = \frac{\frac{r+\overline{\kappa}+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2}}{r+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2}}(\delta - y) + \frac{\overline{\kappa}\lambda\mu_{b2}}{r+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2}}p_2 - \overline{w}_1 + \frac{\lambda\mu_{b1}}{r+\overline{\kappa}+\lambda\mu_{b1}}\left[rp_1 - \delta + y + \overline{\kappa}(p_1 - C_1)\right]}{\frac{r(r+\overline{\kappa}+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2})+\overline{\kappa}\lambda\mu_{b2}}{r+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2}}}.$$

Plugging into (D.15), and using (C.14), (C.23), and the indifference condition which now is $V_{s1} = 0$, we find

$$\overline{w}_1 = \frac{\frac{\lambda\mu_{b1}}{r+\lambda\mu_{b1}}(rp_1 - \delta + y) - \frac{r+\overline{\kappa}+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2}}{r+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2}}(rp_2 - \delta + y)}{1 + \frac{r(r+\overline{\kappa}+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2})+\overline{\kappa}\lambda\mu_{b2}}{\lambda\mu_{s2}(r+\nu\mu_{\overline{\ell}1}+\lambda\mu_{b2})}}.$$

For small search frictions, this equation becomes

$$\overline{w}_1 = \frac{r(p_1 - p_2) - \frac{r\overline{x}}{\lambda\hat{m}_{b1}} - \frac{\overline{\kappa}\overline{x}}{\nu S+\lambda\hat{m}_{b2}}}{1 + \frac{r(nS+\hat{m}_{b2})+\overline{\kappa}\hat{m}_{b2}}{\hat{g}_{s2}(nS+\hat{m}_{b2})}},$$

and is inconsistent with (D.11) since $w_1 < \overline{w}_1$.

Buy asset 1, short asset 2

We consider a "relaxed" problem where asset 1 can be bought instantly and asset 2 can be borrowed instantly at a lending fee of zero. Clearly, if the arbitrage strategy is unprofitable in the relaxed problem, it is also unprofitable when more frictions are present.

Suppose that the arbitrage strategy is profitable. Because the payoff of the strategy is increasing in asset 1's lending fee, there exists a fee $\overline{w}_1 < w_1$ for which the arbitrageur is indifferent between following the strategy and holding no position. When following the strategy, the arbitrageur is always in a state where he holds asset 1 and has borrowed asset 2, because these can be done instantly. If the arbitrageur has not sold asset 2, he can be in four possible states:

- Seeking to lend asset 1. State $\ell 1s2$ with utility $V_{\ell 1s2}$.

- Lent asset 1 to an agent $\underline{s}1$. State $n\underline{s}1s2$ with utility $V_{n\underline{s}1s2}$.
- Lent asset 1 to an agent $\underline{n}1$. State $n\underline{n}1s2$ with utility $V_{n\underline{n}1s2}$.
- Lent asset 1 to an agent $\underline{b}1$. State $n\underline{b}1s2$ with utility $V_{n\underline{b}1s2}$.

If the arbitrageur has sold asset 2, he can be in the four corresponding states that we denote with $n2$ instead of $s2$.

For brevity, we skip the eight flow-value equations, but note that they have a simple solution. To each outcome concerning asset 1 ($\ell 1$, $n\underline{s}1$, $n\underline{n}1$, $n\underline{b}1$) and to each outcome concerning asset 2 ($s2$, $n2$), we can associate a separate utility that we denote by $\hat{V}$. We can then write the utility of a state (which is a "joint" outcome) as the sum of the two separate utilities. For example, the utility $V_{\ell 1s2}$ is equal to $\hat{V}_{\ell 1} + \hat{V}_{s2}$. This decomposition is possible because the outcomes concerning each asset evolve independently.

The utilities $\hat{V}_{\ell 1}$, $\hat{V}_{n\underline{s}1}$, $\hat{V}_{n\underline{n}1}$, and $\hat{V}_{n\underline{b}1}$ are characterized by the flow-value equations

$$\begin{aligned}
r\hat{V}_{\ell 1} &= \nu\mu_{\underline{b}o}(\hat{V}_{n\underline{s}1} - \hat{V}_{\ell 1}) \\
r\hat{V}_{n\underline{s}1} &= \overline{w}_1 + \lambda\mu_{b1}(\hat{V}_{n\underline{n}1} - \hat{V}_{n\underline{s}1}) \\
r\hat{V}_{n\underline{n}1} &= \overline{w}_1 + \underline{\kappa}(\hat{V}_{n\underline{b}1} - \hat{V}_{n\underline{n}1}) \\
r\hat{V}_{n\underline{b}1} &= \overline{w}_1 + \lambda\mu_{s1}(\hat{V}_{\ell 1} - \hat{V}_{n\underline{b}1}).
\end{aligned}$$

and the utilities $\hat{V}_{s2}$, $\hat{V}_{n2}$ are characterized by

$$\begin{aligned}
r\hat{V}_{s2} &= \delta - y + \lambda\mu_{b2}(\hat{V}_{n2} + p_2 - \hat{V}_{s2}) \\
r\hat{V}_{n2} &= \overline{\kappa}(\hat{V}_{s2} - C_2 - \hat{V}_{n2}).
\end{aligned}$$

Solving these equations, we find

$$\begin{aligned}
rV_{\ell 1s2} &= r\hat{V}_{\ell 1} + r\hat{V}_{s2} \\
&= \frac{\frac{\nu\mu_{\underline{b}o}}{r+\nu\mu_{\underline{b}o}}\left(1 - \frac{\lambda\mu_{b1}}{r+\lambda\mu_{b1}}\frac{\underline{\kappa}}{r+\underline{\kappa}}\frac{\lambda\mu_{s1}}{r+\lambda\mu_{s1}}\right)}{1 - \frac{\nu\mu_{\underline{b}o}}{r+\nu\mu_{\underline{b}o}}\frac{\lambda\mu_{b1}}{r+\lambda\mu_{b1}}\frac{\underline{\kappa}}{r+\underline{\kappa}}\frac{\lambda\mu_{s1}}{r+\lambda\mu_{s1}}}\overline{w}_1 + \left[\delta - y + \frac{\lambda\mu_{b2}}{r+\overline{\kappa}+\lambda\mu_{b2}}\left[rp_2 - \delta + y + \overline{\kappa}(p_2 - C_2)\right]\right].
\end{aligned}$$

The arbitrageur is indifferent between initiating the strategy and holding no position if $V_{\ell 1 s 2}$ is equal to p_1. Using this condition, and substituting C_1 from (C.14) and (C.23)

$$\frac{\frac{\nu\mu_{\underline{bo}}}{r+\nu\mu_{\underline{bo}}}\left(1-\frac{\lambda\mu_{b1}}{r+\lambda\mu_{b1}}\frac{\underline{\kappa}}{r+\underline{\kappa}}\frac{\lambda\mu_{s1}}{r+\lambda\mu_{s1}}\right)}{1-\frac{\nu\mu_{\underline{bo}}}{r+\nu\mu_{\underline{bo}}}\frac{\lambda\mu_{b1}}{r+\lambda\mu_{b1}}\frac{\underline{\kappa}}{r+\underline{\kappa}}\frac{\lambda\mu_{s1}}{r+\lambda\mu_{s1}}}\overline{w}_1 = rp_1-\delta+y-\frac{\lambda\mu_{b2}}{r+\lambda\mu_{b2}}(rp_2-\delta+y).$$

For small search frictions, this equation becomes

$$\frac{\hat{g}_{\underline{bo}}}{r+\underline{\kappa}\frac{\hat{g}_{s1}}{r+\underline{\kappa}+\hat{g}_{s1}}+\hat{g}_{\underline{bo}}}\overline{w}_1 = r(p_1-p_2)+\frac{r\overline{x}}{\lambda\hat{m}_{b2}},$$

and is inconsistent with (20) since $w_1 > \overline{w}_1$.

Equations (19) and (20) are jointly satisfied

The two equations are jointly satisfied if

$$\frac{\hat{g}_{\underline{bo}}}{r+\underline{\kappa}\frac{\hat{g}_{s1}}{r+\underline{\kappa}+\hat{g}_{s1}}+\hat{g}_{\underline{bo}}}\frac{w_1}{r} < p_1-p_2 < \frac{w_1}{r}.$$

Substituting p_1 and p_2 from (15) and (16), we can write this equation as

$$A_1\frac{w_1}{r} < \frac{B}{\lambda}+A_2\frac{w_1}{r} < \frac{w_1}{r}, \tag{D.17}$$

where

$$A_2 \equiv \frac{\hat{g}_{\underline{bo}}}{r+\overline{\kappa}+\underline{\kappa}\frac{\hat{g}_{s1}}{r+\overline{\kappa}+\underline{\kappa}+\hat{g}_{s1}}+\hat{g}_{\underline{bo}}} < A_1 \equiv \frac{\hat{g}_{\underline{bo}}}{r+\underline{\kappa}\frac{\hat{g}_{s1}}{r+\underline{\kappa}+\hat{g}_{s1}}+\hat{g}_{\underline{bo}}} < 1$$

and

$$B \equiv \frac{(\phi r+\overline{\kappa})}{(1-\phi)}\left[\frac{1}{\hat{m}_{b2}}-\frac{1}{\hat{m}_{b1}}\right]\frac{\overline{x}}{r} > 0.$$

Equation (D.17) is satisfied if

$$\frac{B}{A_1-A_2} > \frac{\lambda w_1}{r} > \frac{B}{1-A_2}.$$

In this inequality, n enters only through the product λw_1. Therefore, the inequality is satisfied for n in some interval (n_1, n_2). ∎

Proof of Proposition 11: Generalizing the analysis of Section B.2, we can show that a solution for $\varepsilon = 0$ exists, and is close to that for small ε. The limiting equations are (B.23)-(B.26), but with the asset supplies depending on i. For the asymmetric equilibrium, (B.31)-(B.35) generalize to

$$\hat{g}_{\underline{bo}} = \frac{(\overline{\kappa}+\underline{\kappa})\underline{F}}{\underline{\kappa} S_1} \tag{D.18}$$

$$\hat{g}_{s1} = \frac{\overline{\kappa} S_1 + \frac{\overline{\kappa}+\underline{\kappa}}{\underline{\kappa}}\underline{F}}{\hat{m}_{b1}} \tag{D.19}$$

$$\hat{g}_{s2} = \frac{\overline{\kappa} S_2}{\hat{m}_{\overline{b}}}, \tag{D.20}$$

$$\hat{m}_{b1} = \frac{\overline{F}}{\overline{\kappa}} - \sum_{i=1}^{2} S_i - \frac{\underline{F}}{\underline{\kappa}} \tag{D.21}$$

$$\hat{m}_{\overline{b}} = \frac{\overline{F} - \overline{\kappa} S_2}{\overline{F} - \overline{\kappa} S_2 + \underline{F}} \hat{m}_{b1} \tag{D.22}$$

Result (i): An equilibrium where $\nu_1 = \nu$ and $\nu_2 = 0$ can exist if $\Sigma_1 > 0$ and $\Sigma_2 < 0$. Condition $\Sigma_1 > 0$ can be ensured by (11). For small search frictions, condition $\Sigma_2 < 0$ is equivalent to $\hat{g}_{s1} > \hat{g}_{s2}$, as shown in the proof of Proposition 7. Using (D.19), (D.20) and (D.22), we can write condition $\hat{g}_{s1} > \hat{g}_{s2}$ as

$$\left[\overline{\kappa}(S_1 - S_2) + \frac{\overline{\kappa}+\underline{\kappa}}{\underline{\kappa}}\underline{F}\right](\overline{F} - \overline{\kappa} S_2) > \overline{\kappa} S_2 \underline{F}. \tag{D.23}$$

This equation holds for all values of $S_1 \geq S_2$ because Assumption 2 implies that $\overline{F} - \overline{\kappa} S_2 > \overline{\kappa} S_1 \geq \overline{\kappa} S_2$.

Result (ii): The existence condition is now (D.23), but with S_1 and S_2 reversed. It does not hold, for example, when S_1 is large enough to make the term in square brackets negative.

Result (iii): We proceed by contradiction, assuming that for a given $S_1 - S_2 > 0$ there exists an equilibrium where $\nu_1 = \nu_2 = \nu$, even when search frictions converge to zero. Since the parameters a_i, b_i, c_i, and g_i in (C.22) converge to finite limits, while f_i converges to ∞, Σ_i must converge to zero, and $f_i \Sigma_i$ to a finite limit. But then (C.22) implies that the limits of c_1 and c_2 must be the

same. This, in turn, implies that $g_{s1} = g_{s2} \equiv g_s$, which from (B.25) and (B.26) means that

$$\frac{\overline{\kappa} S_i + g_{\underline{bo}} S_i}{m_{\overline{b}} + \frac{\underline{\kappa} g_{\underline{bo}} S_i}{(\overline{\kappa}+\underline{\kappa})(\overline{\kappa}+g_s)}}$$

is independent of i, a contradiction when asset supplies differ. ∎

Proof of Proposition 12: The expected search time for buying asset i is $1/(\lambda \mu_{si})$ and for selling asset i is $1/(\lambda \mu_{bi})$. Thus, our liquidity measure is $\lambda^2 \mu_{bi} \mu_{si} = \lambda(\mu_{bi} \gamma_{si})$. Dropping the multiplicative constant λ and assuming small search frictions, this is equal to $\Lambda_i \equiv \hat{m}_{bi} \hat{g}_{si}$. Equations (D.19) and (D.20) imply that

$$\Lambda_1 = \overline{\kappa} S_1 + \frac{\overline{\kappa} + \underline{\kappa}}{\underline{\kappa}} \underline{F} \tag{D.24}$$

$$\Lambda_2 = \overline{\kappa} S_2. \tag{D.25}$$

Equations (15)-(18), generalized to the case where asset supplies depend on i, imply that the lending fee is

$$w_1 = \theta \left(r + \overline{\kappa} + \underline{\kappa} \frac{\hat{g}_{s1}}{r + \overline{\kappa} + \underline{\kappa} + \hat{g}_{s1}} + \hat{g}_{\underline{bo}} \right) \frac{\underline{x} - \frac{r+\overline{\kappa}+\underline{\kappa}+\hat{g}_{s1}}{r+\overline{\kappa}+\hat{g}_{s1}} (2y - \overline{x})}{\nu(1-\theta) S_1} \tag{D.26}$$

and the price premium is

$$p_1 - p_2 = \frac{(\phi r + \overline{\kappa})}{\lambda(1-\phi)} \left[\frac{1}{\hat{m}_{b2}} - \frac{1}{\hat{m}_{b1}} \right] \frac{\overline{x}}{r} + \theta \hat{g}_{\underline{bo}} \frac{\underline{x} - \frac{r+\overline{\kappa}+\underline{\kappa}+\hat{g}_{s1}}{r+\overline{\kappa}+\hat{g}_{s1}} (2y - \overline{x})}{\nu(1-\theta) S_1 r}. \tag{D.27}$$

Result (i): An increase in $\underline{F}$ increases Λ_1 by (D.24) and leaves Λ_2 constant by (D.25). It increases $\hat{g}_{\underline{bo}}$ by (D.18), decreases $\hat{m}_{b1}$ by (D.21), increases $\hat{m}_{b1}/\hat{m}_{b2} (= \hat{m}_{b1}/\hat{m}_{\overline{b}})$ by (D.22), and increases $\hat{g}_{s1}$ by (D.19). Equation (D.26) then implies that w_1 increases, and Equation (D.27) implies that $p_1 - p_2$ increases. For small search frictions w_1/p_1 varies in the same direction as w_1 since p_1 is close to the limit $(\delta + \overline{x} - y)/r$ while w_1 is close to zero.

Result (ii): A decrease in S_1 decreases Λ_1 by (D.24) and leaves Λ_2 constant by (D.25). Numerical calculations indicate that w_1 and $p_1 - p_2$ increase if $S_1 = S_2 = 0.5$, $\overline{F} = 3$, $\underline{F} = 5.7$, $\overline{\kappa} = 1$, $\underline{\kappa} = 3$, $\phi = \theta = 0.5$, $r = 4\%$, $\delta = 1$, $\overline{x} = 0.4$, $\underline{x} = 1.6$, $y = 0.5$, $\nu/\lambda = 0.25$. If, however, S_1 and S_2 are changed to 1.3, and $\underline{F}$ to 1, while other parameters stay the same, then w_1 and $p_1 - p_2$ decrease. ∎

E The CARA Setting and Miscellanea

This Appendix contains the CARA-based foundation of our model, and a complete proof of Proposition 4. Because this material is not as important as the other proofs, it is written as a separate supplement.

References

Admati, A. R., and P. Pfleiderer (1988): "A Theory of Intraday Patterns: Volume and Price Variability," *Review of Financial Studies*, 1, 3–40.

Aiyagari, R., and M. Gertler (1991): "Asset Returns with Transaction Costs and Uninsurable Individual Risks: A Stage III Exercise," *Journal of Monetary Economics*, 27, 309–331.

Aiyagari, R. S., N. Wallace, and R. Wright (1996): "Coexistence of Money and Interest-Bearing Securities," *Journal of Monetary Economics*, 37, 397–419.

Amihud, Y., and H. Mendelson (1986): "Asset Pricing and the Bid-Ask Spread," *Journal of Financial Economics*, 17, 223–249.

——— (1991): "Liquidity, Maturity, and the Yield on U.S. Treasury Securities," *Journal of Finance*, 46, 479–486.

Barclay, M. J., T. Hendershott, and K. Kotz (2006): "Automation versus Intermediation: Evidence from Treasuries Going Off the Run," *Forthcoming, Journal of Finance*.

Boudoukh, J., and R. F. Whitelaw (1991): "The Benchmark Effect in the Japanese Government Bond Market," *Journal of Fixed Income*, 2, 52–59.

——— (1993): "Liquidity as a Choice Variable: A Lesson from the Japanese Government Bond Market," *Review of Financial Studies*, 6, 265–292.

Buraschi, A., and D. Menini (2002): "Liquidity Risk and Specialness," *Journal of Financial Economics*, 64, 243–284.

Burdett, K., and M. O'Hara (1987): "Building Blocks: An Introduction to Block Trading," *Journal of Banking and Finance*, 11, 193–212.

Chowdhry, B., and V. Nanda (1991): "Multimarket Trading and Market Liquidity," *Review of Financial Studies*, 4, 483–511.

Constantinides, G. M. (1986): "Capital Market Equilibrium with Transaction Costs," *Journal of Political Economy*, 94, 842–862.

Cornell, B., and A. C. Shapiro (1989): "The Misspricing of US Treasury Bonds: a Case Study," *Review of Financial Studies*, 2, 297–310.

Diamond, P. A. (1982): "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, 90, 881–894.

Duffie, D. (1996): "Special Repo Rates," *Journal of Finance*, 51, 493–526.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2002): "Securities Lending, Shorting, and Pricing," *Journal of Financial Economics*, 66, 307–339.

——— (2005): "Over-the-Counter Markets," *Forthcoming, Econometrica*.

——— (2006): "Valuation in Over-the-Counter Markets," Working Paper, Graduate School of Business, Stanford University.

Duffie, D., and Y. Sun (2004): "Existence of Independent Random Matching," Working Paper, Graduate School of Business, Stanford University.

Dupont, D., and B. Sack (1999): "The Treasury Securities Market: Overview and Recent Developments," *Federal Reserve Bulletin*, December, 785–806.

Economides, N., and A. Siow (1988): "The Division of Markets is Limited by the Extent of Liquidity," *American Economic Review*, 78, 108–121.

Ellison, G., and D. Fudenberg (2003): "Knife Edge of Plateau: When do Markets Tip," *Quaterly Journal of Economics*, 118, 1249–1278.

Fleming, M. J. (1997): "The Round-the-Clock Market for U.S. Treasury Securities," *Federal Reserve Bank of New York Economic Policy Review*, pp. 9–32.

——— (2002): "Are Larger Treasury Issues More Liquid? Evidence from Bill Reopenings," *Journal of Money, Credit, and Banking*, 3, 707–35.

——— (2003): "Measuring Treasury Market Liquidity," *Federal Reserve Bank of New York Economic Policy Review*, pp. 83–107.

Goldreich, D., B. Hanke, and P. Nath (2002): "The Price of Future Liquidity: Time-Varying Liquidity in the U.S. Treasury Market," Working Paper, Institute of Finance and Accounting, London Business School.

Graveline, J. J., and M. R. McBrady (2004): "Who Makes the On-The-Run Treasuries Special?," Working Paper, Graduate School of Business, Stanford University.

Heaton, J., and D. J. Lucas (1996): "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing," *Journal of Political Economy*, 104, 443–487.

Huang, M. (2003): "Liquidity Shocks and Equilibrium Liquidity Premia," *Journal of Economic Theory*, 109, 104–129.

Ibbotson (2004): *Stock, Bonds, Bills, and Inflation Statistical Yearbook*. Ibbotson Associates, Chicago.

Jordan, B. D., and S. D. Jordan (1997): "Special Repo Rates: An Empirical Analysis," *Journal of Finance*, 52, 2051–2072.

Keim, D. B., and A. Madhavan (1996): "The Upstairs Market for Large-Block Transactions: Analysis and Measurement of Price Effects," *Review of Financial Studies*, 9, 1–36.

Kiyotaki, N., and R. Wright (1989): "On Money as a Medium of Exchange," *Journal of Political Economy*, 97, 927–954.

Krishnamurthy, A. (2002): "The Bond/Old-Bond Spread," *Journal of Financial Economics*, 66, 463–506.

Lo, A. W., H. Mamaysky, and J. Wang (2004): "Asset Prices and Trading Volume under Fixed Transactions Costs," *Journal of Political Economy*, 112, 1054–1090.

Mason, R. (1987): "The 10-year Bond Markets," Credit Suisse First Boston, CSFB Research.

Moulton, P. C. (2004): "Relative Repo Specialness in U.S. Treasuries," *Journal of Fixed Income*, 14, 40–49.

Pagano, M. (1989): "Endogenous Market Thinnes and Stock Price Volatility," *Review of Economic Studies*, 269-287.

Strebulaev, I. (2002): "Liquidity and Asset Pricing: Evidence from the U.S. Treasury Securities Market," Working Paper, Graduate School of Business, Stanford University.

Sundaresan, S. (2002): *Fixed Income Markets and Their Derivatives*. South-Western Publishing Company.

Trejos, A., and R. Wright (1995): "Search, Bargaining, Money, and Prices," *Journal of Political Economy*, 103(1), 118–141.

Vayanos, D. (1998): "Transaction Costs and Asset Prices: A Dynamic Equilibrium Model," *Review of Financial Studies*, 11, 1–58.

Vayanos, D., and J.-L. Vila (1999): "Equilibrium Interest Rate and Liquidity Premium with Transaction Costs," *Economic Theory*, 13, 509–539.

Vayanos, D., and T. Wang (2006): "Search and Endogenous Concentration of Liquidity in Asset Markets," Working Paper, London School of Economics.

Wallace, N. (2000): "A Model of the Liquidity Yield Structure Based on Asset Indivisibility," *Journal of Monetary Economics*, 45, 55–68.

Warga, A. (1992): "Bond Returns, Liquidity, and Missing Data," *Journal of Financial and Quantitative Analysis*, 27, 605–617.

Weill, P.-O. (2004): "Liquidity Premia in Dynamic Bargaining Markets," Working Paper, Finance Department, NYU Stern School of Business.