The Strategic Determination of the Supply of Liquid Assets

Athanasios Geromichalos
University of California – Davis

Lucas Herrenbrueck
Simon Fraser University

Revised version: October 2020

ABSTRACT

We study the joint determination of asset supply and asset liquidity in a model where financial assets can be liquidated for money in over-the-counter (OTC) secondary markets. Traders choose to enter the market where they expect to find the best terms; understanding this, asset issuers choose their quantities strategically in order to profit from the liquidity services their assets confer. We find that small differences in OTC microstructure can induce very large differences in the relative liquidity of two assets. Our model has a number of applications, including the superior liquidity of U.S. Treasuries over equally safe corporate debt.

JEL Classification: E31, E43, E52, G12

Keywords: monetary-search models, OTC markets, endogenous liquidity, endogenous asset supply

Contact: ageromich@ucdavis.edu, herrenbrueck@sfu.ca

We would like to thank Aleksander Berentsen, Darrell Duffie, Tai-Wei Hu, Robert Jones, Oscar Jorda, Ricardo Lagos, Ed Nosal, Guillaume Rocheteau, Ina Simonovska, Alan Taylor, Dimitri Vayanos, Venky Venkateswaran, Pierre-Olivier Weill, and Randall Wright for their useful comments and suggestions, as well as participants at the 2015 St Louis FED workshop on Money, Banking, Payments, and Finance, the WEAI 90th Annual Conference, the 12th Annual Macroeconomic Workshop in Vienna, Austria, the Spring 2016 Midwest Macro Meetings, the Fall 2017 Midwest Macro Meetings, the 11th NYU Search Theory Workshop, the 2018 North American Summer Meeting of the Econometric Society, and at seminars at the University of Wisconsin, Madison, the University of Saskatchewan, and the University of British Columbia. We also thank Zijian Wang and Sukjoon Lee for their invaluable help.

1 Introduction

Why do U.S. Treasuries sell at higher prices than corporate or municipal bonds with similar characteristics, even after controlling for safety?[1] A popular answer is "due to their *liquidity*". More precisely, the Treasury sells its bonds at a *premium* because investors expect to be able to (re)sell these bonds easily in the secondary market and are, thus, willing to pay higher prices in the primary market.[2] While this is a plausible explanation, some important questions remain. Why are the secondary markets for other types of bonds less liquid than the one for Treasuries? Is it hard(er) for sellers to find buyers due to some hardwired market friction (e.g., a poorly organized interdealer network)? Or, are there not enough buyers drawn to those markets to whom I could sell my bonds – and if so, why? Or, perhaps finding trading partners is not so hard, but there are not enough bonds to go around in the market? Finally, how do these candidate explanations (and their interaction) affect asset prices and liquidity in general equilibrium?

To answer these questions, we develop a model of the joint determination of the supply of potentially liquid assets and their realized liquidity. Key to our model is the fact that this liquidity does not only depend on the (exogenous) characteristics of the market an asset trades in, but also on the (endogenous) decision of agents to visit that market. Our model has three main ingredients. The first is an empirically relevant concept of asset liquidity: agents can liquidate assets for money in Over-the-Counter (OTC) secondary markets which, as in Duffie, Gârleanu, and Pedersen (2005), are characterized by search and bargaining. This implies that assets are imperfect substitutes for money and have, generally, positive liquidity premia. The second ingredient of our model is an entry decision by the agents. Each asset trades in a distinct OTC market, and agents choose to visit the market where they expect to find the best terms. The third ingredient is strategic interaction among asset issuers: the agencies that issue assets realize that equilibrium asset prices – thus, the rate at which they can borrow – depend not only on their own decisions but also on those made by issuers of similar (hence, competing) assets. Specifically, we focus on two issuers of assets who play a differentiated Cournot game, where, crucially, the product (asset) differentiation stems from differences in the *microstructure* of the secondary market where each asset trades.

As a starting point, we study the endogenous determination of OTC market participation, keeping asset supplies fixed. This can be seen as the subgame following the supply decisions of issuers, and it provides a number of new insights on its own. Agents receive an idiosyncratic shock that determines whether they will need, ex-post, additional liquidity in the secondary market (i.e., sell assets) or whether they will be the providers of that liquidity (i.e., buy assets).

[1] For a thorough discussion of this stylized fact, see Krishnamurthy and Vissing-Jorgensen (2012).

[2] For instance, former Assistant Secretary of the U.S. Treasury, Brian Roseboro, points precisely in this direction: "A deep, liquid, and resilient secondary market serves our goal of lowest-cost financing for the taxpayer by encouraging more aggressive bidding in the primary market." (*A Review of Treasury's Debt Management Policy*, June 3, 2002, available at http://www.treas.gov/press/releases/po3149.htm.)

An agent who turns out to be an asset seller can only visit one OTC market at a time; since, typically, assets are costly to own due to the liquidity premium, agents choose to 'specialize' ex-ante in asset A or B. Unlike sellers, who must take into account the cost of holding a particular asset, the asset buyers make their market choice in a more 'elastic' way since their money is good to buy any asset. As a result, when one of the markets, say market A, has any kind of advantage – an exogenous matching advantage or simply offering bigger surpluses because there are more A-assets to be traded – asset buyers rush into that market more eagerly than sellers. In turn, this implies that the trade probability in that market for sellers increases by far more than that for buyers. Crucially, it is the sell-probability that affects the issue price, because someone who buys an asset (in the primary market) cares about the ease of selling it later. Through this channel, small differences in market microstructure can be *magnified* into a big endogenous liquidity advantage for one asset, even with constant returns to scale (CRS) in the OTC matching technology. And with a modest degree of increasing returns to scale (IRS), demand curves can be *upward sloping*, because an asset in large supply is likely to be more liquid.

Next, we study the duopoly game between the issuers, who realize that the outcome of the subgame determines the demand for their assets. When the matching technology exhibits CRS, asset supplies tend to be strategic substitutes. In this case, equilibrium issue sizes are low, and the prices of both assets include liquidity premia. When the matching technology exhibits IRS, asset supplies tend to be strategic complements. This promotes aggressive competition among issuers, in the sense that equilibrium issue sizes can be large, and that equilibria of the subgame tend to be in a corner in which only one of the two OTC markets operates. Effectively, one of the assets ends up illiquid. Therefore, our paper does not only endogenize the supply of (potentially) liquid assets, but also their degree of liquidity; this is precisely why we have been careful about reminding the reader that assets are 'potentially' liquid.

We also study how changes in the exogenous market microstructure affect equilibrium play, and, consequently, asset prices and liquidity premia. More precisely, letting δ_i, $i = A, B$, denote the matching efficiency in the OTC market for asset i, we fix δ_A and study the effect of changes in δ_B. Suppose the matching process is CRS and δ_B falls slightly below δ_A; in this case, issuer A increases her asset supply and issuer B decreases it, but the strategic pattern of a Cournot game is maintained. The exogenous liquidity advantage of asset A is magnified by the entry choices of agents, which, in turn, feeds back into a rising (falling) liquidity premium on asset A (B). As δ_B declines further, there comes a point at which issuer A has an incentive to boost up her supply and drive B out of the secondary market altogether. At that point asset B becomes fully illiquid. As δ_B falls even further, the threat of competition by asset B becomes so insignificant that issuer A practically turns into a monopolist in the supply of liquid assets.

With a degree of IRS in the matching technology, this process is accelerated. For a reasonable parametrization of the model, we show that asset B will become completely illiquid even if the matching function in market B is almost equally efficient as the one in market A (say,

$\delta_B = 0.99\,\delta_A$), and there is only a tiny amount of IRS in the matching function.[3] If one were to look at these numbers, one might infer that asset B cannot be much less liquid than asset A. This conclusion would be mistaken, because it would be based only on the exogenous factors. What is more important is that agents endogenously choose to concentrate their trade in market A because they expect other agents will do the same – and, reinforcing this, because both issuers have an incentive to compete for this concentration by issuing large (enough) amounts.

The model has a number of fruitful applications. The first is the superior liquidity of U.S. Treasuries over *equally safe* corporate or municipal bonds. One may argue that this stylized fact has an easy explanation: the secondary market for Treasuries is more well-organized (which in our model would be captured by a more efficient matching technology). However, the relative illiquidity of corporate or municipal bonds has been well-documented for many decades. If the key behind this illiquidity was just some poorly organized secondary markets, one wonders why the issuers of these bonds have not taken steps to improve the efficiency of these markets, which would lower the rate at which they can borrow. Hence, it seems unlikely that the stylized fact in question can be purely explained by differences in market efficiency. Our model can offer a deeper explanation: perhaps Treasuries have a small exogenous advantage over other types of bonds, but this is amplified into a large endogenous liquidity advantage by the fact that investors choose to concentrate their trade into the secondary market for Treasuries, rather than get exposed to the *liquidity risk* associated with trading other types of bonds.[4]

Our model can also shed some light on the well-documented empirical observation that for many types of bonds, there is a positive relationship between bond supply and the realized liquidity premium (see Hotchkiss and Jostova, 2007, and Alexander, Edwards, and Ferri, 2000, for the case of corporate bonds). This relationship also seems to be embraced by practitioners: common advice given to first-time sovereign bond issuers is that "the issue should be large enough to assure market liquidity" (Das, Polan, and Papaioannou, 2008), and in May 1998, the U.S. Treasury announced that it would discontinue issuance of 3-year notes and reduce the issuance frequency of 5-year notes from monthly to quarterly "in order to continue to assure large, liquid issues" (quote reproduced from Fleming, 2002). As we have already seen, our model suggests that with even a slight degree of IRS in matching, an increase in the supply of an asset can lead to a higher liquidity premium.

Furthermore, our model can help explain how consolidating secondary markets would be beneficial for asset issuers, a belief commonly held among practitioners. In a recent report on

[3] Specifically, the elasticity of the number of matches to scale (i.e., the total number of entrants) only has to be 1.02 or larger. For context, a scale elasticity of 1 is CRS, and most theoretical finance papers use a congestion-free matching function with scale elasticity 2. That is, our model can be a mix of 98% CRS and 2% congestion-free matching, yet outcomes *look like* those obtained with the latter – precisely because competition between the issuers makes it so.

[4] For instance, Oehmke and Zawadowski (2016) and Helwege and Wang (2016) report that many investors choose to not participate in the corporate bonds markets altogether, because they are highly concerned about the risk of not being able to liquidate their bonds quickly and at good terms, if such a need arises.

the corporate bond market structure (BlackRock, 2014), the authors make a number of proposals that they believe could increase the "deteriorating" liquidity of corporate bonds. One of their main suggestions is that regulators should work towards consolidating the secondary markets for corporate bonds. This view is supported by the empirical findings of Oehmke and Zawadowski (2016), who find that "the *fragmented* nature of the corporate bond market impedes its liquidity" (emphasis added). While these papers touch upon some features of corporate bonds that our model abstracts from (such as "standardization"), they are clearly implying that merging secondary markets would improve the bonds' liquidity. Our theory predicts precisely that. Specifically, with even a slight degree of IRS in matching, a merging of secondary markets would increase the liquidity premia enjoyed by the issuers (in the primary market), because the market consolidation reduces the investors' risk of not being able to sell.

Finally, our model delivers some important results regarding welfare. First, and most importantly, there exists no monotonic relationship between welfare and "liquidity" (for any measure of liquidity we could choose). Second, unlike output, social welfare tends to be maximized for small-to-intermediate quantities of liquid assets. This alone does not tell us whether a monopoly or a Cournot duopoly of liquid assets would be superior; each is possible, depending on parameters. However, it does tell us that aggressive competition for secondary market liquidity, where issuers issue large amounts and drive liquidity premia to zero, is suboptimal. Consequently, market segmentation and exogenous liquidity differences can be good for welfare because they tend to discourage such aggressive competition.

The present paper is related to a branch of the recent literature, often referred to as "New Monetarism" (see Lagos, Rocheteau, and Wright, 2017), that has highlighted the importance of asset liquidity for the determination of asset prices. See for example Geromichalos, Licari, and Suárez-Lledó (2007), Lagos and Rocheteau (2008), Lester, Postlewaite, and Wright (2012), Nosal and Rocheteau (2013), Andolfatto and Martin (2013), Andolfatto, Berentsen, and Waller (2013), and Hu and Rocheteau (2015). In these papers assets are 'liquid' because they serve as a medium of exchange in frictional decentralized markets.[5] In some other papers, liquidity properties stem from the fact that assets serve as collateral, as in Venkateswaran and Wright (2013) and Andolfatto, Martin, and Zhang (2015).[6] The majority of this literature has studied asset liquidity (and prices) under the simplifying assumption that asset supply is fixed. Recent exceptions include Rocheteau and Rodriguez-Lopez (2014) and Branch, Petrosky-Nadeau, and Rocheteau (2016). Moreover, Bethune, Sultanum, and Trachter (2017) consider an environment with asset issuance and decentralized secondary markets, but they focus on efficiency and pol-

[5] Consequently, in most of these papers, assets compete with money as media of exchange. In recent work, Fernández-Villaverde and Sanches (2016) extend the Lagos and Wright (2005) framework to study the interesting question of competition among privately issued electronic currencies, such as Bitcoin and Ethereum.

[6] Some papers within this literature have shown that adopting models where assets are priced both for their role as stores of value and for their liquidity may be the key to rationalizing certain asset pricing-related puzzles. See Lagos (2010), Geromichalos, Herrenbrueck, and Salyer (2016), and Herrenbrueck (2019b).

icy rather than liquidity. Our paper is also related to Caramp (2017) who endogenizes asset creation with a focus on asset quality and asymmetric information.

A key difference of our paper with the works mentioned so far is that here asset liquidity is *indirect*. Assets never serve as media of exchange (or as collateral) to purchase consumption. Their liquidity stems from the fact that agents can sell them for money in a secondary market. This idea is exploited in a number of recent papers, including Geromichalos and Herrenbrueck (2016), Berentsen, Huber, and Marchesiani (2014, 2016), Herrenbrueck (2019a), Mattesini and Nosal (2016), and Geromichalos, Herrenbrueck, and Lee (2018). As argued earlier, we believe that this approach is empirically relevant for a large class of financial assets. A common feature of these papers is that a secondary asset market allows agents to rebalance their liquidity after an idiosyncratic expenditure need has been revealed. This idea draws upon the work of Berentsen, Camera, and Waller (2007), where the channeling of liquidity takes place through a competitive banking system. Our work is also related to Lagos and Zhang (2015), but in that paper agents use money to purchase assets (rather than goods) in an OTC financial market.

Our work is also related to the literature initiated by the seminal work of Duffie et al. (2005), which studies how frictions in OTC financial markets affect asset prices and trade. A non-exhaustive list of such papers includes Vayanos and Wang (2007), Weill (2007, 2008), Vayanos and Weill (2008), Lagos and Rocheteau (2009), Lagos, Rocheteau, and Weill (2011), Afonso and Lagos (2015), Üslü (2015), Chang and Zhang (2015). Our paper is uniquely distinguished from all these papers, starting with the very concept of liquidity: we have a monetary model where agents sell assets for cash after learning of a consumption opportunity, while in those papers, agents differ in the utility flow derived from holding an asset and pay for assets with transferable utility. Furthermore, we characterize the strategic incentives facing issuers of potentially liquid assets, and thereby endogenize the supply of such assets in addition to their liquidity.

Our paper is also related to a strand of the Industrial Organization literature that studies the effect of secondary markets for durable goods on the producers' pricing decisions. Examples include Rust (1985, 1986). In these papers, the existence of a secondary market, where buyers could sell the durable good in the *future*, affects the pricing decisions of sellers *now* through affecting the buyers' willingness to pay for the good.[7] In our model, if secondary markets were shut down (so that assets have to be held to maturity), agents would be only willing to buy assets at their fundamental value. The existence of secondary markets endows assets with (indirect) liquidity properties, which, in turn, allows issuers to borrow funds at lower rates (i.e., sell bonds at a price that includes a liquidity premium).

The paper is organized as follows. Section 2 describes the model. In Section 3, we study the economy with exogenous asset supplies, and in Section 4, we endogenize asset supplies by characterizing the game between asset issuers. Section 5 analyzes a special case of our model

[7] Within the context of financial rather than commodity markets, this idea is also exploited by Geromichalos et al. (2016) and Arseneau, Rappoport, and Vardoulakis (2015).

where assets are perfect substitutes and we can obtain closed-form solutions, and Section 6 concludes. Appendix A.1 discusses empirical counterparts of our modeling choices, and Appendix A.2 contains some technical details of the model. Finally, the Web Appendix contains several extensions of our analysis – only one asset issuer being strategic, one asset issuer being a Stackelberg leader, and one asset issuer having a higher cost of creating assets than the other – and an analytical characterization of the equilibria in our model.

2 The model

Time is discrete and the horizon is infinite. Each period consists of three sub-periods where different economic activities take place. In the first sub-period, two distinct OTC financial markets open, denoted by OTC_j, $j = \{A, B\}$. Agents who hold assets of type j can sell them for money in OTC_j. One could think of asset A as T-Bills and asset B as corporate AAA bonds. In the second sub-period, agents visit a decentralized goods market where trade is bilateral, and agents are anonymous and lack commitment. We refer to this market as the DM. Due to the aforementioned frictions, trade necessitates a medium of exchange in the DM, and this role can be played only by money. During the third sub-period, economic activity takes place in a centralized market, which is similar in spirit to the settlement market of Lagos and Wright (2005) (henceforth, LW). We refer to this market as the CM. There are two permanently distinct types of agents, buyers and sellers, named by their role in the DM, and the measure of both types of agents is normalized to the unit. Agents live forever. There are also two agencies, $j = \{A, B\}$, that issue asset j in its respective *primary market* which opens within the third sub-period.

All agents discount the future between periods (but not sub-periods) at rate $\beta \in (0, 1)$. Buyers consume in the DM and CM sub-periods and supply labor in the CM sub-period. Their preferences within a period are given by $\mathcal{U}(X, H, q) = X - H + u(q)$, where X, H represent consumption and labor in the CM, respectively, and q consumption in the DM. Sellers consume only in the CM, and they produce in both the CM and the DM. Their preferences are given by $\mathcal{V}(X, H, h) = X - H - q$, where X, H are as above, and q stands for units of production in the DM. We assume that u is twice continuously differentiable with $u' > 0$, $u'(0) = \infty$, $u'(\infty) = 0$, and $u'' < 0$. Let q^* denote the optimal level of production in a bilateral meeting in the DM, i.e., $q^* \equiv \{q : u'(q^*) = 1\}$. The issuers of assets are only present in the CM. Their preferences are given by $\mathcal{Y}(X, H) = X - H$, where X, H are as above. The issuers also discount the future at rate β. What makes them special is that they can issue assets that potentially carry liquidity premia, thus allowing them to obtain net profits out of this operation.[8]

[8] Alternatively, one could assume that the issuers have to finance certain expenditures and, hence, have to borrow at least a certain amount, but can choose to borrow more if doing so is profitable. As long as that lower bound is not too large, our results would remain valid under the alternative specification.

We now provide a detailed description of the various sub-periods. In the third sub-period, all agents consume and produce a general good or fruit. All agents (including the issuers) have access to a technology that transforms one unit of labor into one unit of the fruit. Agents can choose to hold any amount of money which they can purchase at the ongoing price φ_t (in real terms). The supply of money is controlled by the monetary authority, and it evolves according to $M_{t+1} = (1+\mu)M_t$, with $\mu > \beta - 1$. New money is introduced, or withdrawn if $\mu < 0$, via lump-sum transfers to buyers in the CM. Money has no intrinsic value, but it possesses all the properties that make it an acceptable medium of exchange in the DM (e.g., it is portable, storable, and recognizable by agents). Agents can also purchase any amount of asset j at price p_j, $j = \{A, B\}$ (in nominal terms). These assets are one-period nominal bonds: each unit of (either) asset purchased in period t's CM pays one dollar in the CM of $t+1$.[9] Let the supply of the assets be denoted by (A_t, B_t). In Section 3, we will treat them as fixed; in Section 4, they will be chosen strategically by the issuers. Each issuer chooses the supply of her asset as a best response to her rival's action in order to maximize profits, realizing that both her own and her rival's assets provide indirect liquidity services to an asset purchaser.

After making their portfolio decisions in the CM, buyers receive an idiosyncratic consumption shock: a measure $\ell < 1$ of buyers will have a desire to consume in the forthcoming DM. We refer to these buyers as the C-types, and to the remaining $1-\ell$ buyers as the N-types ("not consuming"). Since buyers did not know their type when they made their portfolio choices, N-types will typically hold some cash that they will not use in the current period, while C-types may find themselves short of cash (since carrying money is costly). The OTC round of trade is placed *after* the idiosyncratic uncertainty has been resolved, but *before* the DM opens to allow a reallocation of money into the hands of those who value it most. OTC financial markets are segmented: an agent who wants to sell or purchase assets is free to enter either OTC_A or OTC_B, but she must choose one market at a time.[10] Hence, coordination is extremely important, and agents will pick the market where they expect to find better trading conditions.

Once C-types and N-types have decided which market they wish to enter, a matching function, $f_j(C_j, N_j)$, brings together sellers (C-types) and buyers (N-types) of assets in the OTC_j, in bilateral matches. Throughout the paper we use the specific functional form:

$$f_j(x,y) = \delta_j \left(\frac{xy}{x+y}\right)^{1-\rho} (xy)^{\rho},$$

with $\delta_j \in [0,1]$ and $\rho \in [0,1]$, and thus $f_j(x,y) \leq \min\{x,y\}$. The term δ_j captures exogenous efficiency factors in OTC_j, such as the density of the dealer network. The term $\rho \in [0,1]$ governs returns to scale in matching; for concreteness, notice that the elasticity of the number

[9] Since the assets are nominal, in steady state their supply must grow at rate μ, too (see, for example, Berentsen and Waller, 2011).

[10] We discuss and justify this assumption in Appendix A.1. Furthermore, perfectly integrated markets are equivalent to one special case of our model, explored in Section 5.

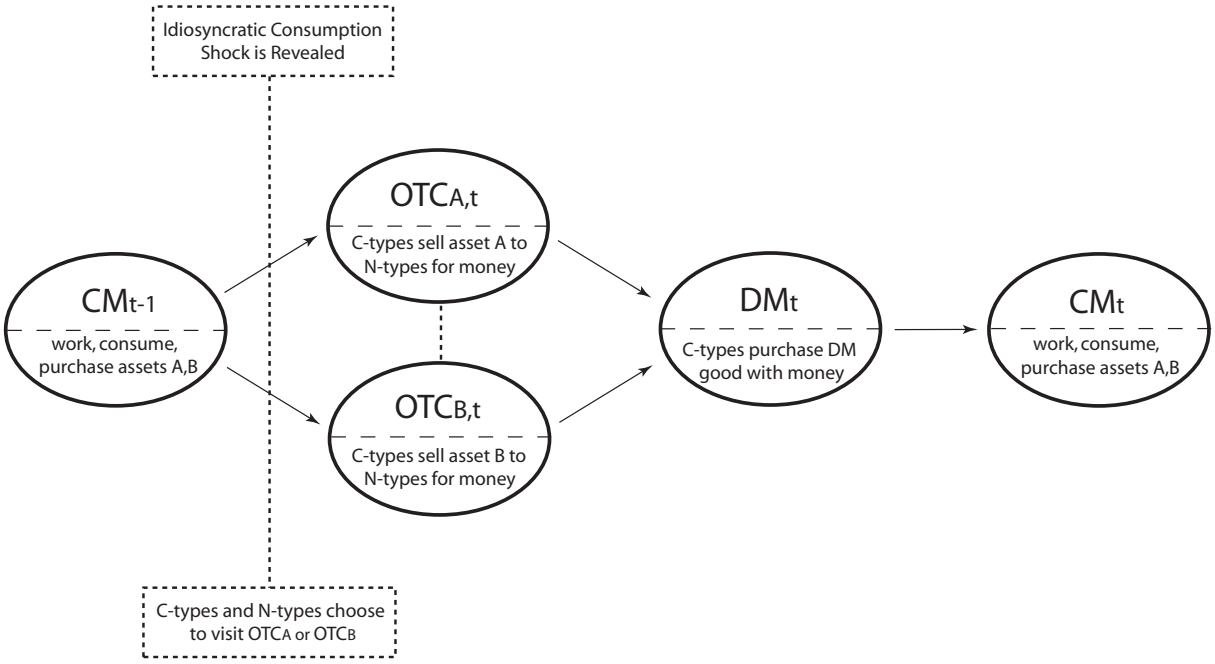


Figure 1: Timing of events.

of matches with respect to scale, keeping the ratio of buyers to sellers fixed, is $1+\rho$. This functional form allows us to study both the case of CRS ($\rho = 0$, scale elasticity of 1) and IRS ($\rho > 0$, scale elasticity exceeding 1). Within any match in either of the OTC markets the C-type makes a take-it-or-leave-it (TIOLI) offer with probability $\theta \in (0,1)$, otherwise the N-type does.

The second sub-period is the standard decentralized goods market of the LW model. C-type buyers meet bilaterally with sellers and negotiate over the terms of trade. Exchange must take place in a *quid pro quo* fashion, and only money can serve as a medium of exchange.[11] Since all the interesting insights of the paper follow from agents' interaction in the OTC round of trade, we wish to keep the DM as simple as possible. To that end, we assume that all C-type buyers match with a seller, and that in any match the buyer makes a TIOLI offer to the seller.

Figure 1 summarizes the timing of the main actions in the model. It is important to highlight that the secondary OTC markets are completely separate from the primary markets where assets are first issued. Nevertheless, the microstructure of the secondary markets, summarized by the parameters $\delta_j, \rho,$ and θ, will determine the liquidity properties of the assets and, consequently, their selling price in the primary market.

[11] Here we shall make this an assumption of the model. However, a number of recent papers in the monetary-search literature, such as Rocheteau (2011) and Lester et al. (2012) do not place any restrictions on which objects can serve as media of exchange and show that, under asymmetric information, fiat money will endogenously arise as a superior medium of exchange, thus, providing a micro-founded justification for our assumption.

3 The economy with exogenous asset supply

In this section we analyze the economy, treating the supplies of assets (A, B) as given. The task of endogenizing the asset supplies is carried out in Section 4.

In order to streamline the analysis, we relegate the details of defining the value functions and characterizing the terms of trade in the OTC markets and the DM to Appendix A.2. Here we include a summary. All agents in the economy have linear preferences over labor and consumption goods in the CM, which will induce linear value functions in the CM, and make a number of economic decisions easy to characterize. First, consider a DM meeting between a seller and a C-type buyer who brings a quantity m of money. The buyer will either buy the first-best quantity q^*, or, if her money is not enough, spend all of it on the quantity $q = \varphi m < q^*$. Second, consider a meeting in the OTC market for asset $j \in \{A, B\}$, where the N-type brings a quantity $\tilde{m}$ of money, and the C-type brings a portfolio (m, d_j) of money and asset j. With probability $1-\theta$, the N-type makes a TIOLI offer, in which case she buys the C-type's assets and compensates him with the least amount of money that the C-type will accept. With probability θ, the C-type makes a TIOLI offer, in which case he receives the N-type's money and compensates her with assets valued at par.[12]

What is the probability of matching in an OTC market for an individual agent? First, let $e_{_C} \in [0, 1]$ and $e_{_N} \in [0, 1]$ denote the fractions of C-types and N-types, respectively, who choose to enter OTC_A. Then, the measure of asset sellers and buyers in OTC_A is given by $e_{_C}\ell$ and $e_{_N}(1-\ell)$, respectively, and the measure of asset sellers and buyers in OTC_B is given by $(1-e_{_C})\ell$ and $(1-e_{_N})(1-\ell)$. Letting $\alpha_{ij} \in [0, 1]$ denote the matching probabilities for agents of type $i = \{C, N\}$ in OTC_j, $j = \{A, B\}$, we have:

$$\alpha_{CA} \equiv \frac{f_A\big(e_{_C}\ell, e_{_N}(1-\ell)\big)}{e_{_C}\ell}, \qquad \alpha_{CB} \equiv \frac{f_B\big((1-e_{_C})\ell, (1-e_{_N})(1-\ell)\big)}{(1-e_{_C})\ell}, \tag{1}$$

$$\alpha_{NA} \equiv \frac{f_A\big(e_{_C}\ell, e_{_N}(1-\ell)\big)}{e_{_N}(1-\ell)}, \qquad \alpha_{NB} \equiv \frac{f_B\big((1-e_{_C})\ell, (1-e_{_N})(1-\ell)\big)}{(1-e_{_N})(1-\ell)}. \tag{2}$$

[12] In OTC trade, three kinds of outcomes are possible: (a) the C-type's asset holdings could limit the trade; (b) the N-type's money holdings could limit the trade; (c) or both are so large that the pooled money is enough to purchase the first-best DM quantity ($m + \tilde{m} > q^*/\varphi$), and the C-type has enough assets to compensate the N-type. In Geromichalos and Herrenbrueck (2016), we showed that assets can only be priced (in the CM) at a determinate liquidity premium if case (a) applies in the corresponding OTC market. Case (c) is also relevant as the boundary of case (a), where an asset becomes abundant and the liquidity premium converges to zero. Case (b), however, only complicates the general equilibrium analysis. Since it does not feature a positive liquidity premium, and since our interest is in asset issuers who seek to exploit such a premium, we exclude case (b) from our analysis. This is done by assuming that inflation is not too large, so that all agents carry at least *half* of the first-best amount of money.

3.1 Optimal behavior

As shown in Appendix A.2, the sellers' decisions in this model are trivial. To that end, in what follows we will use the term 'agents' to refer to buyers, i.e., the agents who consume in the DM and make interesting portfolio decisions. In the OTC market, these agents will take on roles as 'asset sellers' and 'asset buyers' depending on the outcome of their consumption shock (C or N, respectively). The 'seller'-agents who produce and sell goods in the DM will not come up again in the main text.

As is standard in models that build on LW, all agents choose their optimal portfolio independently of their trading histories in preceding markets. This result follows from the "no-wealth-effects" property, which, in turn, stems from the quasilinear preferences. What is new here is that in addition to choosing an optimal portfolio of money and assets, $(\hat{m}, \hat{d}_A, \hat{d}_B)$, agents also choose which OTC market they will enter in order to sell or buy assets once their type has been revealed. The agent's choice can be analyzed with an objective function, denoted by $J(\hat{m}, \hat{d}_A, \hat{d}_B)$, which summarizes the cost and benefit from choosing portfolio $(\hat{m}, \hat{d}_A, \hat{d}_B)$. To obtain J, substitute the values of trading in the OTC markets and in the DM (Equations A.4-A.8, derived in the appendix) into the maximization operator of the CM value function (Equation A.1). After using the linearity of the value function itself (Equation A.2), we can drop all terms that do not depend on the choice variables $(\hat{m}, \hat{d}_A, \hat{d}_B)$ to obtain the objective function:

$$J(\hat{m}, \hat{d}_A, \hat{d}_B) = -\varphi\left(\hat{m} + p_A\hat{d}_A + p_B\hat{d}_B\right) + \beta\hat{\varphi}\left(\hat{m} + \hat{d}_A + \hat{d}_B\right) + \beta\ell\left[u\left(\hat{\varphi}\hat{m}\right) - \hat{\varphi}\hat{m} + \max\left\{\underbrace{\theta\alpha_{CA}S_{CA}}_{\text{enter A}}, \underbrace{\theta\alpha_{CB}S_{CB}}_{\text{enter B}}\right\}\right], \tag{3}$$

so that the optimal portfolio choice is fully described by $\max J$, where the current prices of money and assets, (φ, p_A, p_B), and the future price of money, $\hat{\varphi}$, are taken as given.

The interpretation of the objective function is intuitive. The first term represents the cost that the agent needs to pay in order to purchase the portfolio $(\hat{m}, \hat{d}_A, \hat{d}_B)$ in the CM, and the second term represents the benefit from selling these assets in the CM of the next period. If one were to shut down the DM market (say, by setting $\ell = 0$), there would be no liquidity considerations and the agent's objective function would consist only of these two terms. The third term reveals that with probability ℓ the agent will be a C-type in the next period. In this case she can use her money $(\hat{m})$ to purchase consumption in the DM (generating a net surplus equal to $u\left[\hat{\varphi}\hat{m}\right] - \hat{\varphi}\hat{m}$), and she can enter OTC$_j$, $j = A, B$, in order to acquire more money by selling her assets ($\hat{d}_A$ or $\hat{d}_B$). In the last expression, the terms S_{Cj} represent the surplus for the C-type in OTC$_j$, but the agent will actually enjoy this surplus only if she gets to match in that market *and* make

the TIOLI offer, an event that occurs with probability $\theta\alpha_{Cj}$.[13] Exploiting the OTC bargaining solution (i.e., Lemma A.2) and Equation (A.9), one can verify that, for $j = \{A, B\}$:

$$S_{Cj} = \begin{cases} u(q^*) - u(\hat{\varphi}\hat{m}) - q^* + \hat{\varphi}\hat{m}, & \text{if } \hat{d}_j > m^* - \hat{m}, \\ u(\hat{\varphi}\hat{m} + \hat{\varphi}\hat{d}_j) - u(\hat{\varphi}\hat{m}) - \hat{\varphi}\hat{d}_j, & \text{otherwise,} \end{cases} \tag{4}$$

where the condition $\hat{d}_j > m^* - \hat{m}$ states that in this case the agent's asset holdings are "abundant", i.e., they allow her to reach the first-best amount of money, m^*, through OTC trade.

Two important observations are in order. First, while we have only imposed an exogenous segmentation assumption on the OTC markets, an *endogenous* segmentation will arise in the *primary* markets: i.e., agents will typically choose to purchase only asset A or asset B in the CM. In equilibrium, assets will trade at a premium, and agents will only pay this premium if they expect to sell the asset in the OTC. Since they can only enter one OTC (and anticipate having to choose eventually), they will choose *ex-ante* (i.e., in the CM), to "specialize" in asset A or B.[14] This, in turn, implies that an agent's portfolio choice is intertwined with the choice of which OTC market to enter in case she turns out to be a C-type. For instance, we shall see that agents who choose to trade in a less liquid OTC market will self-insure against the liquidity shock by carrying more money.

The second important observation is that the agent's choice of which market to enter if she turns out to be an N-type is unrelated with her choice of asset specialization in the CM. This is because the N-type's asset and money holdings do not affect the bargaining solution in OTC trade (see Lemma A.2). As a result, regardless of her asset choice which by the time the N-type makes her OTC entry choice is sunk, this agent will enter OTC_A only if:

$$(1-\theta)\alpha_{NA}S_{NA} \geq (1-\theta)\alpha_{NB}S_{NB}.$$

In the last expression, the terms S_{Nj} represent the surplus for the N-type in OTC_j. Exploiting Lemma A.2 and equation (A.10), one can verify that, for $j = \{A, B\}$,

$$S_{Nj} = \begin{cases} u(q^*) - u(\hat{\varphi}\tilde{m}) - q^* + \hat{\varphi}\tilde{m}, & \text{if } \tilde{d}_j > [u(q^*) - u(\hat{\varphi}\tilde{m})]/\hat{\varphi}\,, \\ \hat{\varphi}\tilde{m} + \hat{\varphi}\tilde{d}_j - u^{-1}\left[u(\hat{\varphi}\tilde{m}) + \hat{\varphi}\tilde{d}_j\right], & \text{otherwise,} \end{cases} \tag{5}$$

[13] One may wonder why there is no $(1-\ell)$-term in the objective function. Does the N-type not generate value by bringing money into the OTC? Yes, this is the case, as the full value function (Equation A.1) shows. But the technical restriction (6), justified in Footnote 12, guarantees that the N-type's money is never *marginal* in OTC trade. Hence the N-branch can be dropped from the portfolio choice problem; the only decision to be made along the N-branch is which OTC market to enter.

[14] Agents may still hold the other asset if indifferent, i.e., if that asset is abundant or illiquid.

where $(\tilde{m}, \tilde{d}_j)$ stand for the N-type's expectation about the money and asset-j holdings, respectively, that her trading partner, a C-type, will carry into OTC$_j$. The condition $\tilde{d}_j > [u(q^*) - u(\hat{\varphi}\tilde{m})]/\hat{\varphi}$ states that the asset holdings of the C-type are large enough to allow her post-OTC money balances to reach the first-best amount, m^*.

3.2 Equilibrium

In steady state, the cost of holding money can be summarized by the parameter $i \equiv (1+\mu-\beta)/\beta$; exploiting the Fisher equation, this parameter represents the nominal interest rate on an *illiquid* asset. For example, in any equilibrium it must be true that $p_j \geq 1/(1+i)$, $j = \{A, B\}$, since otherwise there would be an infinite demand for the assets; however, the inequality could be strict if the assets are liquid. The restriction $\mu > \beta - 1$ translates into $i > 0$. We also assume that:

$$i < \ell(1-\theta)\left[u'\left(q^*/2\right) - 1\right], \tag{6}$$

a technical restriction. It ensures that $q_{0j} > q^*/2$ for every agent, thus the N-type's money will never be the limiting factor in OTC trade. See our explanation in Footnote 12, and note that if we did have $q_{0j} < q^*/2$, the implied burden of the inflation tax would be enormous.

We have thirteen endogenous variables.[15] First, we have the equilibrium real balances $\{z_A, z_B\}$ held by the agent who chooses to specialize in asset A or B (recall from the discussion in Section 3.1 that an agent who chooses to trade in OTC$_A$ will typically make different portfolio choices than one who chooses to trade in OTC$_B$).

Next, we have the equilibrium quantities $\{q_{0A}, q_{1A}, q_{0B}, q_{1B}, \tilde{q}_{1A}, \tilde{q}_{1B}\}$. The first four represent the quantity of DM good purchased by a C-type agent who either did not trade in the OTC market (indexed by 0), or who traded *and* made the TIOLI offer (indexed by 1), depending on whether they chose to specialize in asset A or asset B. The last two terms (i.e., the $\tilde{q}$'s) represent the quantity of DM good purchased by an agent who traded in her chosen OTC, A or B, but did not get to make the TIOLI offer. The purchasing power of the C-type in the DM will depend on whether she got to make the offer or not, and, naturally, we have $q_{1j} \geq \tilde{q}_{1j}$, for all j.[16]

Next, we have the prices of the three assets $\{\varphi, p_A, p_B\}$. Finally, we have the entry choices $\{e_C, e_N\}$, i.e., the fractions of C-types and N-types, respectively, who choose to enter OTC$_A$.

We now show that seven out the thirteen endogenous variables can be derived from the following six variables, $\{q_{0A}, q_{1A}, q_{0B}, q_{1B}, e_C, e_N\}$. First, we have $z_j = q_{0j}$, for $j = \{A, B\}$, since the C-type who does not trade in the OTC can only purchase the amount of DM goods that her

[15] This count excludes the terms of trade in the OTC markets, since they follow directly from the main endogenous variables described in this section and Lemma A.2.

[16] More precisely, we have $q_{1j} > \tilde{q}_{1j}$, unless the C-type's asset holdings satisfy $d_j \geq [u(q^*) - u(\varphi m)]/\varphi$. Then, even if the N-type makes the offer the C-type can afford a money transfer of $m^* - m$, and we have $q_{1j} = \tilde{q}_{1j} = q^*$.

own real money holdings, z_j, allow her to afford. Second, the price of money solves:

$$\varphi M = e_{_C} q_{0A} + (1 - e_{_C}) q_{0B}. \tag{7}$$

This equation is the market clearing condition in the market for money. Third, the equilibrium asset prices must satisfy the demand equations:[17]

$$p_j = \frac{1}{1+i}\Big(1 + \ell\, \alpha_{Cj}\theta \cdot [u'(q_{1j}) - 1]\Big), \quad \text{for} \quad j = \{A, B\}. \tag{8}$$

For future reference, notice that as long as $q_{1j} < q^*$, the marginal unit of the asset allows the agent to acquire additional money which she can use in order to boost her consumption in the DM. In this case, the agent is willing to pay a *liquidity premium* in order to hold the asset. On the other hand, if $q_{1j} = q^*$, the term inside the square brackets becomes zero, and $p_j = 1/(1+i)$, which is simply the *fundamental price* of a one-period nominal bond.

Finally, the quantities consumed in the DM by agents who did not make the TIOLI offer in the preceding OTC market satisfy:

$$\tilde{q}_{1A} = \min\left\{q^*,\ u^{-1}\left(u(q_{0A}) + \varphi\,\frac{A}{e_{_C}}\right)\right\}, \tag{9}$$

$$\tilde{q}_{1B} = \min\left\{q^*,\ u^{-1}\left(u(q_{0B}) + \varphi\,\frac{B}{1-e_{_C}}\right)\right\}, \tag{10}$$

where φ has been explicitly defined as a function of the variables q_{0j} in (7). (These equations are derived from substituting equilibrium variables into part (b) of Lemma A.2.)

The analysis so far establishes that if one had solved for $\{q_{0A}, q_{1A}, q_{0B}, q_{1B}, e_{_C}, e_{_N}\}$, then the remaining seven variables could also be immediately determined. Hence, hereafter we refer to these six variables as the "core" variables of the model. We now turn to the description of the equilibrium conditions that determine the core variables. Throughout this discussion, recall that the terms $e_{_C}, e_{_N}$ are also implicitly affecting the arrival rates α_{Cj}.

First, the money demand equation for those specializing in asset j:

$$i = \ell\ (1 - \alpha_{Cj}\theta) \cdot [u'(q_{0j}) - 1] + \ell\, \alpha_{Cj}\theta \cdot [u'(q_{1j}) - 1], \quad \text{for} \quad j = \{A, B\}. \tag{11}$$

Note that we have defined $\alpha_{ij} = 0$ if there is no entry at all into market j. If that is the case, q_{0j} and q_{1j} are still defined as limits even though nobody actually trades at those quantities.

Next, the OTC trading protocol links q_{0j} and q_{1j}. Consider for instance market A. The bargaining solution, evaluated at equilibrium quantities, becomes:

[17] These follow directly from obtaining the first-order conditions in the agent's objective function, i.e., Equation (3), and imposing equilibrium quantities. Notice that the asset prices do not only depend on the variables q_{1j}, but also on the equilibrium values of $e_{_C}, e_{_N}$ which affect the arrival rates α_{Cj}; see Equations (1).

$$q_{1A} = \min\left\{q^*,\ q_{0A} + \frac{\varphi A}{e_{_C}}\right\},$$

where $\varphi A/e_{_C}$ is the real value of assets that the C-type brings into OTC_A.[18] Even though the real aggregate supply of asset A is φA, the agent under consideration holds more than the average because some agents do not hold asset A at all (they specialize in asset B). After substituting the price of money from Equation (7) into the last expression, we obtain two equations, one for each market:

$$q_{1A} = \min\left\{q^*,\ q_{0A} + \frac{A}{M}\cdot\frac{e_{_C}q_{0A} + (1-e_{_C})q_{0B}}{e_{_C}}\right\}, \tag{12}$$

$$q_{1B} = \min\left\{q^*,\ q_{0B} + \frac{B}{M}\cdot\frac{e_{_C}q_{0A} + (1-e_{_C})q_{0B}}{1-e_{_C}}\right\}. \tag{13}$$

If it happens that $e_{_C} = 1$ (no C-types enter the B-market) and $B > 0$, then we define $q_{1B} = q^*$ as a limit, because a C-type of infinitesimal size who decided to deviate and hold asset B could hold the entire stock of it, which would certainly satiate them in an OTC trade – in the hypothetical case that there was an N-type in the B-market willing to trade with them. Similarly, if $e_{_C} = 0$ and $A > 0$, then we define $q_{1A} = q^*$.

How large can the aggregate supply of an asset be for the asset to remain scarce in OTC trades? Clearly, the asset is more likely to be scarce if its ownership is *diluted*, i.e., if many agents choose to hold that asset in the CM. So for example, asset A is most likely to be scarce if $e_{_C} = 1$. But in this special case, Equation (12) tells us that the asset is scarce ($q_{1A} < q^*$) only if the condition $1 + A/M < q^*/q_{0A}$ is satisfied. On the boundary, $q_{1A} = q^*$, so we can use the money demand equation (11) to obtain the bounds:

$$\bar{A} \equiv M\left(q^*/\bar{q}_{0A} - 1\right),\ \text{where}\ \bar{q}_{0A}\ \text{solves}\ i = \left[\ell - \theta\, f_A(\ell, 1-\ell)\right]\left[u'(\bar{q}_{0A}) - 1\right],$$
$$\bar{B} \equiv M\left(q^*/\bar{q}_{0B} - 1\right),\ \text{where}\ \bar{q}_{0B}\ \text{solves}\ i = \left[\ell - \theta\, f_B(\ell, 1-\ell)\right]\left[u'(\bar{q}_{0B}) - 1\right].$$

There are three things to notice here. First, if $A > \bar{A}$, then asset A is certain to be abundant but the reverse is not always true, because asset ownership can be *concentrated* in the hands of a few agents. Second, if we did fix $e_{_C} = 1$ so that ownership of asset A was maximally diluted, then asset A would indeed be abundant if and only if $A \geq \bar{A}$, and conversely for asset B. Third, if the market for asset A has an exogenous liquidity advantage ($\delta_A > \delta_B$), then $\bar{A} > \bar{B}$, and vice versa. For convenience, we define the maximal upper bound on asset supply beyond which either asset is certain to be abundant:

[18] If the C-type's asset holdings are plentiful in the OTC, then we know that this agent will be able to purchase the first-best amount of money in the DM, hence, $q_{1A} = q^*$. On the other hand, if the asset is scarce in OTC trade, the C-type gives away all of her assets, $\varphi A/e_{_C}$. Moreover, since here we are in the case where the C-type makes the offer, she will swap assets for money at a one-to-one ratio. As a result, in equilibrium it must be that $q_{1A} = q_{0A} + \varphi A/e_{_C}$, which explains the last expression.

$$\bar{D} \equiv \max\{\bar{A}, \bar{B}\}.$$

The remaining task is to characterize the OTC market entry choices. Consider first a C-type. As we have already discussed, this type at the beginning of the period has already made the choice to hold either asset A or asset B, so the choice of which market to enter has effectively been made. Evaluating equation (4) at equilibrium quantities, we find that if the C-type makes the TIOLI offer, her surplus of trading in market $j \in \{A, B\}$ equals:[19]

$$S_{Cj} = u(q_{1j}) - u(q_{0j}) - q_{1j} + q_{0j}. \tag{14}$$

But since the agent's portfolio choice effectively determines her market choice if she turns out to be a C-type, this surplus has to be balanced not only against the probability of needing to trade, actually matching, and making the offer ($\ell \times \alpha_{Cj} \times \theta$), but also against the cost of carrying the asset. Hence, we define the "net" surplus that the agent obtains if she chooses to specialize in asset j to be:

$$\tilde{S}_{Cj} \equiv -iq_{0j} - [(1+i)p_j - 1](q_{1j} - q_{0j}) + \ell\left[u(q_{0j}) - q_{0j}\right] + \ell\alpha_{Cj}\theta S_{Cj}.$$

We can use the money and asset demand equations (8 and 11) to substitute for i and p_j in the last expression. After some algebra, we obtain:

$$\tilde{S}_{Cj} = \ell\left(1 - \alpha_{Cj}\theta\right) \cdot \left[u(q_{0j}) - u'(q_{0j})q_{0j}\right] + \ell\alpha_{Cj}\theta \cdot \left[u(q_{1j}) - u'(q_{1j})q_{1j}\right]. \tag{15}$$

Thus, in equilibrium, the C-types' portfolio choice $e_{_C}$ must satisfy:

$$e_{_C} \quad = \quad \begin{cases} 1, & \text{if } \tilde{S}_{CA} > \tilde{S}_{CB}, \\ 0, & \text{if } \tilde{S}_{CA} < \tilde{S}_{CB}, \\ \in [0,1], & \text{if } \tilde{S}_{CA} = \tilde{S}_{CB}. \end{cases} \tag{16}$$

Finally, we want to characterize the market choice of the N-type agents. Since these agents are asset buyers their own asset holdings do not matter, so they can enter the market for either asset independently of which asset they chose to hold in the preceding CM. Thus, an N-type will simply enter the market in which she expects a greater surplus, accounting for the probability of trading and making the TIOLI offer. Evaluating equation (5) at equilibrium quantities

[19] This equality holds regardless of whether the asset is plentiful in the OTC meeting or not. Consider first the case of plentiful assets. For this case evaluating the relevant (i.e., the "abundant") branch of Equation (4) at equilibrium quantities yields $S_{Cj} = u(q^*) - u(q_{0j}) - q^* + q_{0j}$, which is exactly what one would obtain if $q_{1j} = q^*$ was imposed on Equation (14). Next, consider the case of scarce assets and for simplicity focus on OTC_A. In this case, evaluating (4) at equilibrium quantities yields $S_{Cj} = u(q_{1j}) - u(q_{0j}) - \varphi A/e_{_C}$, where $\varphi A/e_{_C}$ is the real value of assets that the C-type brings into OTC_A. But as we know from the discussion that leads to Equation (12), here $q_{1A} = q_{0A} + \varphi A/e_{_C}$. Hence, the validity of Equation (14) is once again verified.

implies that the surplus for the N-type who chooses to enter OTC$_A$ is given by:

$$S_{NA} = \begin{cases} u(q^*) - u(q_{0A}) - q^* + q_{0A}, & \text{if } A/e_{_C} > [u(q^*) - u(q_{0A})]/\varphi\,, \\ q_{0A} + \varphi\frac{A}{e_{_C}} - u^{-1}\left(\varphi\frac{A}{e_{_C}} + u(q_{0A})\right), & \text{otherwise,} \end{cases} \tag{17}$$

and the surplus for the N-type who chooses to enter OTC$_B$ is given by:

$$S_{NB} = \begin{cases} u(q^*) - u(q_{0B}) - q^* + q_{0B}, & \text{if } B/(1-e_{_C}) > [u(q^*) - u(q_{0B})]/\varphi\,, \\ q_{0B} + \varphi\frac{B}{1-e_{_C}} - u^{-1}\left(\varphi\frac{B}{1-e_{_C}} + u(q_{0B})\right), & \text{otherwise.} \end{cases} \tag{18}$$

In (17) and (18) we have used the value of money, φ, to keep these expressions relatively short, but it is understood that φ is itself a function of the core variables, defined in (7).

Thus, in equilibrium, the N-types' entry choice $e_{_N}$ must satisfy:

$$e_{_N} = \begin{cases} 1, & \text{if } \alpha_{NA}S_{NA} > \alpha_{NB}S_{NB}, \\ 0, & \text{if } \alpha_{NA}S_{NA} < \alpha_{NB}S_{NB}, \\ \in [0,1], & \text{if } \alpha_{NA}S_{NA} = \alpha_{NB}S_{NB}. \end{cases} \tag{19}$$

We can now define a steady-state equilibrium in the model with fixed asset supplies:

Definition 1. Assume (for now) that asset supplies are fixed and equal to $(A, B) \in \mathbb{R}^2_+$. A steady-state equilibrium for the core variables of the model is a list $\{q_{0A}, q_{1A}, q_{0B}, q_{1B}, e_{_C}, e_{_N}\}$ such that Equations (11) for $j = \{A, B\}$, (12), and (13) hold, and agents' entry choices satisfy Equations (16) and (19).

3.3 Characterization of equilibrium

We are now ready to characterize the equilibria of the economy, summarized by the core variables $\{q_{0A}, q_{1A}, q_{0B}, q_{1B}, e_{_C}, e_{_N}\}$, conditional on the asset supplies $A, B \geq 0$. Before we go to the technical details, it is helpful to gain some intuition by considering the optimal entry decision of the representative N-type, who takes as given the term $e_{_N}$, the proportion of other N-types who enter the A-market, and best responds by entering in either market A or B. A higher value of $e_{_N}$ implies a bigger congestion among N-types in market A, a force that *discourages* our representative N-type from entering into that market. On the other hand, a higher $e_{_N}$ implies that a larger fraction of C-types will be drawn to market A, because C-types like a market with many N-types, and this force *encourages* our representative N-type to enter into that market. And, to make things even more interesting, a higher value of $e_{_C}$ implies that the supply of asset A,

which is fixed for now, will be *diluted* among a larger number of agents (this channel becomes more relevant if the supply of asset A is scarce). Hence, in any bilateral meeting in OTC$_A$, the surplus is more likely to be limited because the C-type is constrained by her asset holdings, yet another force that discourages our representative N-type from entering into market A.

Summing up, an increase in the term $e_{_N}$ generates multiple and opposing forces, and may have non-monotonic effects on the optimal entry decision of the representative N-type. What one can say safely is that everything else equal, the typical N-type is more likely to enter into market A if: (i) $\delta_A > \delta_B$, because then the former market has an exogenous matching advantage; and (ii) $A > B$, because then there is a larger potential surplus when trading asset A.

Moving to the formal analysis, we construct equilibria as fixed points of $e_{_N}$. To be specific: first, we fix a level of $e_{_N}$; then we solve for the optimal portfolio choices through Equations (11)-(13) and (16); and finally, we define the N-types' *reply function*:

$$G(e_{_N}) \equiv \frac{\alpha_{NA} S_{NA} - \alpha_{NB} S_{NB}}{\alpha_{NA} S_{NA} + \alpha_{NB} S_{NB}},$$

where the surplus (S) and match probability (α) terms have the optimal choices substituted. This function measures the relative benefit to an *individual* N-type from choosing the A-market over the B-market, assuming a proportion $e_{_N}$ of all *other* N-type agents enters the A-market, and all other decisions are conditionally optimal. To make it easier to visualize, G is scaled to lie between -1 and +1. A value of $e_{_N}$ is part of an "interior" equilibrium if $e_{_N} \in (0,1)$ and $G(e_{_N}) = 0$, or a "corner" equilibrium if $e_{_N} = 0$ and $G(0) \leq 0$ or $e_{_N} = 1$ and $G(1) \geq 0$.

Proposition 1. *The following types of equilibria exist, and have these properties:*

(a) There exists a corner equilibrium where $e_{_C} = e_{_N} = 0$; only the B-market is open for trade.

(b) There exists a corner equilibrium where $e_{_C} = e_{_N} = 1$; only the A-market is open for trade.

(c) Assume $\rho = 0$ (CRS) and asset supplies are low enough so that assets are scarce in OTC trade. Then, $\lim_{e_N \to 0+} G(e_{_N}) > 0 > G(0)$ and $\lim_{e_N \to 1-} G(e_{_N}) < 0 < G(1)$; the corner equilibria are <u>not</u> robust to small trembles. There exists at least one interior equilibrium which is robust to small trembles.

(d) Assume $\rho > 0$ (IRS). Then, $\lim_{e_N \to 0+} G(e_{_N}) = G(0) < 0$ and $\lim_{e_N \to 1} G(e_{_N}) = G(1) > 0$; the corner equilibria now <u>are</u> robust to small trembles. There exists at least one interior equilibrium, which may or may not be robust to small trembles.

(e) Assume $\rho = 0$ (CRS) and $\delta_A = \delta_B$ (equal market quality). Then, a symmetric equilibrium exists where $e_{_C} = e_{_N} = A/(A+B)$, $q_{0A} = q_{0B}$ and $q_{1A} = q_{1B}$, and $p_A = p_B$.

(f) If, in addition to the assumptions in (e), $A = B < \bar{D}/2$ (asset supplies are equal and small), $i \to 0$ (low inflation), $\theta\delta(1-\ell) < 0.5$ (not-too-high bargaining power for the C-type), and $u''' \geq 0$ (convex marginal utility), then $G'(0.5) < 0$; that is, the symmetric equilibrium is robust to small trembles.

Proof. See Sections C.1-C.3 in the Web Appendix. □

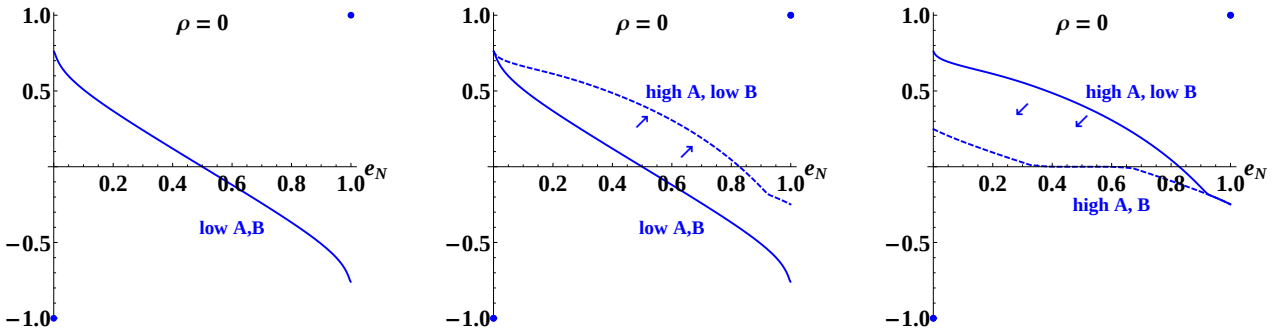


Figure 2: The reply function $G(e_N)$ for CRS ($\rho = 0$) and varying asset supplies.

In all cases, the C-types' entry choice e_C is optimally adjusting in the background, and it is generally an increasing function of e_N; when there are many buyers in a market, sellers would like to go to the same market. Of course, nobody would try to trade in a ghost town, so it must be the case that $e_C = 0$ if and only if $e_N = 0$, and $e_C = 1$ if and only if $e_N = 1$ (parts (a) and (b) of the proposition). Therefore, the corners are always equilibria.

These results are depicted in Figure 2, which shows how the reply function G depends on e_N and on asset supplies, given CRS in matching. Blue dots at $G(0) = -1$ and $G(1) = +1$ indicate that the corners are always equilibria. In the left panel, G is shown for relatively low supplies of A, B, and there is an interior fixed point at $e_N = 0.5$. As shown in part (c) of the Proposition, the corners are not robust to small trembles, but the interior fixed point is: if a few more N-types accidentally enter the A-market, individual N-types have an incentive to deviate back to B. In the middle panel, we show what happens for a higher supply of A: the G-function shifts up and more agents trade in the A-market, but the equilibrium is still robust.

The right panel illustrates the case where both A and B are high: the G-function shifts back down, but now it contains a flat segment for intermediate values of e_N. This is due to the fact that with high asset supplies, the aforementioned *dilution effect* disappears: if the supply of assets is high enough, each individual C-type will be able to achieve q^* in the DM after they sell their assets in the OTC (even as the fixed asset supply gets diluted among more C-types). With the dilution effect out of the picture, a higher e_N implies a higher congestion effect in market A but also a larger measure of C-types in that market (i.e., a higher equilibrium e_C). With CRS in matching these two effects completely offset each other, leading to a flat G-function; or, equivalently, a continuum of equilibria with $e_C = e_N$ when asset supplies are large enough.

We now move on to the case of IRS in the matching technology, corresponding to part (d) of Proposition 1. Figure 3 shows the reply function G under $\rho = 0.5$, an intermediate degree of IRS. In this case, a high value of e_N still implies some congestion among N-types, but this effect is dominated by the large measure of C-types drawn to market A (precisely because e_N is high). Does that mean that G will be strictly increasing? Not necessarily. Consider for instance the left panel of the figure, where both asset supplies are small, so that the dilution effect is active.

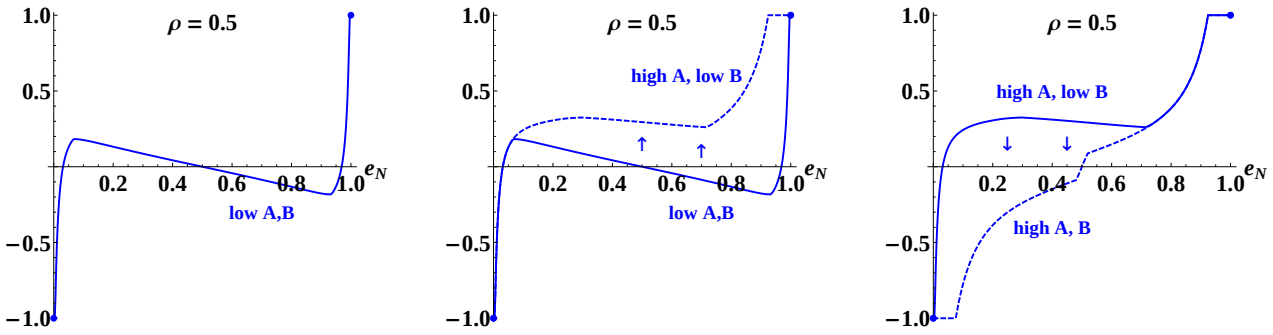


Figure 3: The reply function $G(e_N)$ for IRS ($\rho = 0.5$) and varying asset supplies.

If e_N is large, the typical N-type has a high probability of matching in market A because that market is flooded with C-types (as well as N-types). But each of those C-types is carrying only a tiny fraction of the supply of asset A, which was small to begin with. This force discourages the representative N-type from entering market A, and gives G the non-monotone shape seen in the left panel of Figure 3. More precisely, that picture shows that there are five equilibria: the two corners (which are both robust under small errors now), the robust interior equilibrium, and two non-robust asymmetric equilibria.

What if the supply of asset A was high but that of asset B stayed low? This case is illustrated in the middle panel of Figure 3, where one can see that the robust interior equilibrium is now *eliminated*. Unless trade was concentrated in the B-corner in the first place, N-types now have an incentive to migrate to the A-market, C-types will follow, and ultimately all trade will be in the A-corner. Finally, the right panel depicts the case where both A and B are high. In this case, the G-function shifts down (compared to the high-A, low-B case), and incentives to trade in the B-market are restored. However, when both asset supplies are large, the dilution effect vanishes and the G-function becomes increasing throughout, so the corners are the only robust equilibria. There does exist an interior equilibrium by continuity, but if it was ever played, a small shock would drive the agents into one of the corners.

What is important here is that to obtain this result we do not need increasing returns to be particularly strong. As Figure 3 shows, robust interior equilibria can exist even under increasing returns – but only if asset supplies are small enough. (For a formal analysis of this case, see Section C.3 in the Web Appendix.) This is why accounting for the endogenous choice of asset supply by issuers is so important, a task we will carry out in Section 4. We shall see there that even with $\rho \ll 1$, the outcome of the issuers' game *looks as extreme as* the outcome with $\rho = 1$, because competition between issuers makes it so.

Finally, as part (e) of Proposition 1 shows, the system admits a simple symmetric solution in one special case which we call "balanced CRS": there are CRS in OTC market matching ($\rho = 0$) and neither asset has an exogenous liquidity advantage ($\delta_A = \delta_B$).[20] Such equilibria can be

[20] We use the word "balanced" to describe the assumption $\delta_A = \delta_B$. We could also call it "symmetric", but

solved in closed form with a judicious choice of utility function, as we explore in Section 5 below. And as part (f) of the Proposition shows, with a few more technical assumptions we can prove that $G(e_N)$ is downward-sloping in a neighborhood of the symmetric equilibrium (as depicted in the left panel of Figure 2); thus, this particular interior equilibrium is robust.

Beyond the results of Proposition 1, a general analytical characterization is not possible and most of the analysis which follows will be numerical. (The model can also not be simplified without losing essential insights.[21]) Throughout the rest of the paper, we maintain the parameters $u(q) = \log(q)$, $\ell = 0.5$, $\theta = 0.5$, $i = 0.1$, and $M = 1$, which yield $\bar{A} = 0.8/(4 - \delta_A)$ and $\bar{B} = 0.8/(4 - \delta_B)$. In the rest of Section 3, we vary the asset supplies A and B exogenously, and in Section 4, they will be chosen by strategic issuers. Throughout, we vary the parameters of OTC microstructure (δ_A, δ_B, ρ).

For given parameters, we guess a starting point for e_N, then iterate the function $G(e_N)$ in the direction of its sign, until convergence or until reaching a corner. Specifically, we use $e_N^0 \equiv \delta_A A/(\delta_A A + \delta_B B)$ as an efficient starting point for iteration; if a robust interior equilibrium exists, it is likely to involve more entry into the market with a higher matching probability, and/or higher trading volume. If the corners are not robust, this procedure will always find an interior equilibrium. On the other hand, a robust interior equilibrium may exist but not be found if a corner is robust and the starting point is close to it.

3.4 Comparative statics

Now that we understand the structure of possible equilibria, we want to compare asset prices in these equilibria, and interpret the comparative statics of prices with respect to quantities as the aggregate demand for these assets. These comparative statics are shown in Figure 4. In all graphs, the supply of asset A is on the horizontal axis and the supply of B is held fixed and

we reserve that word for equilibria where *all* variables indexed by A equal their B-counterparts (e.g., $p_A = p_B$). Even in the balanced environment, there are asymmetric equilibria: the corner equilibria for one, and additional asymmetric interior equilibria if $\rho > 0$, as shown in the left panel of Figure 3.

[21] We have a core system of six equations, and most of the endogenous variables show up in multiple equations. Moreover, the equations are non-linear and include kinks, due to the various branches that characterize the agents' market entry decisions. One may wonder whether some simplifying assumptions would allow us to achieve a complete analytical characterization. We believe that the model presented here constitutes the most parsimonious framework that can capture all the salient features of the question we are studying, hence, any further simplification would eliminate insights that we think are essential. A few examples may clarify this point. A simplifying assumption often adopted in these types of models is that the bargaining power of agents is equal to either 0 or 1. (This is precisely what we assume for the DM, because not many interesting things happen in that market.) Imposing such an assumption in the OTC would be a bad idea: it would imply that either the C-types or the N-types get no surplus from OTC trade, which would render their entry decision indeterminate. As we have explained, the agent's decision about which market to visit is one of the most important economic forces in our model. As another example, some papers (e.g., Mattesini and Nosal, 2016) gain tractability by assuming that asset trade takes place *only* in OTC markets, and the original asset holdings are given to agents in the CM as endowments, i.e., there is no primary asset market. Clearly, such an assumption here would deprive the model of one of its most important ingredients, the endogenous determination of asset supply.

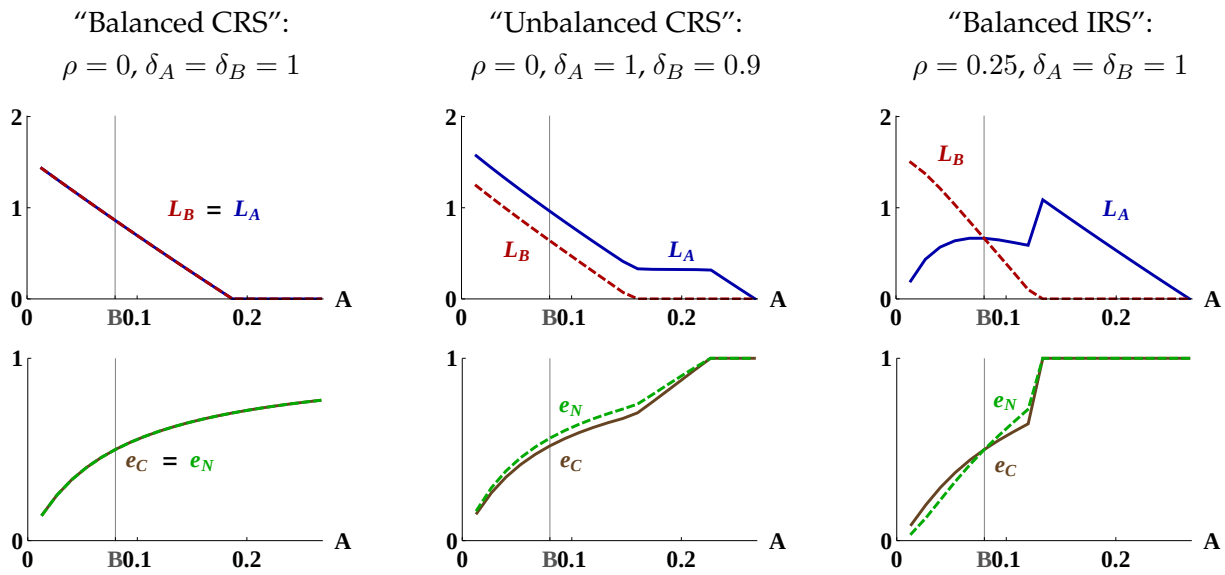


Figure 4: Net liquidity premia L_j (in %) and entry choices, varying A and holding B fixed (indicated by a vertical line).

indicated by a gray vertical line. We show three cases: first, the simplest case of balanced CRS ($\rho = 0$ and $\delta_A = \delta_B$); second, giving an exogenous advantage to asset A ($\delta_A > \delta_B$); and third, without an advantage for either asset but with IRS in matching ($\rho > 0$). In all three examples, the graphs in the top row show the *net liquidity premia* of assets A and B, defined as:

$$L_j \equiv (1+i)p_j - 1 = \ell\, \alpha_{Cj}\theta \left[u'(q_{1j}) - 1\right].$$

The graphs in the bottom row of the figure show the market entry choices e_C and e_N.

Notice first that some standard results are replicated in our model. First, the liquidity premium of an asset is zero if that asset is in very large supply, no matter how liquid the market for that asset is. The reason is that as the asset supply becomes large enough, $q_{1j} \to q^*$, and thus, $u'(q_{1j}) \to 1$. (One should be careful with terms here: the asset does not "lose" its liquidity properties in this case, they only become inframarginal. The asset still contributes to the overall supply of liquidity in the sense that money demand will be lower than it would be if that asset did not exist.) Furthermore, real balances decrease with inflation so the need to liquidate assets in the OTC markets becomes stronger with inflation; if the asset supplies are small enough, the liquidity premium on any liquid asset will rise with inflation, too.

In addition to these standard results, our model also delivers new insights into asset pricing in this environment of segmented OTC markets. Three results stand out. The first is that when matching in the markets satisfies "balanced CRS" (that is, CRS and neither market having an exogenous liquidity advantage), there exists a unique interior equilibrium when the asset supplies are not too large. In this equilibrium, $e_C = e_N = A/(A+B)$, so the ratio of buyers to sellers

is 1 in each market, we have $p_A = p_B$, and all the equilibrium quantities and prices only depend on the *sum* of the asset supplies, $A + B$. Thus, the assets turn out to be perfect substitutes in general equilibrium even though their secondary markets are completely segmented. Part (e) of Proposition 1 shows this formally, and the leftmost column of Figure 4 illustrates it.

The second result from this section is that exogenous liquidity differences are *amplified* by the market entry process, even with CRS. Consider a case where $\delta_A > \delta_B$, so that OTC$_A$ has an exogenous liquidity advantage. As illustrated in the middle column of Figure 4, both e_C and e_N increase, but the latter increases more. Intuitively, the N-types only consider the potential trading surplus in the OTC market when deciding which market to enter, while the C-types also consider the ex-ante cost of carrying either asset, and therefore the N-types are more sensitive to liquidity differences when choosing their market. The end result is that market tightness from the point of view of asset sellers rises in the more liquid market and falls in the less liquid one: formally, we observe that the elasticity of the endogenous ratio α_{CA}/α_{CB} with respect to the exogenous ratio δ_A/δ_B is bigger than 1. Crucially, it is the point of view of OTC asset *sellers* that matters for asset pricing at the issue stage; people who buy a newly issued asset are concerned about the conditions at which they can sell it down the road, but people who plan to buy the asset later in the secondary market have no influence on the issue price. As a consequence, even a small divergence of δ_A and δ_B will drive a larger wedge between the liquidity premia on the two assets. We view this result as Step 1 of an explanation why two assets with otherwise similar features can have big differences in their liquidity – most prominently, of course, U.S. Treasuries compared to equally safe corporate or municipal bonds.

The third result from this section is that IRS in matching encourage market concentration, i.e., corner equilibria. This is illustrated in the rightmost column of Figure 4. Near the origin, we have a case of $A \ll B$, so asset A is barely traded in OTC markets (though not entirely absent due to the fact that ownership of asset B is much more diluted). As the supply of A increases, more agents are willing to trade it in the OTC market because of the increase in potential trading surplus; and crucially, N-types are more sensitive to this increase, so the ratio e_N/e_C rises as A increases. This is important because again, it means that asset A becomes rapidly more attractive to C-types through *two* channels (market tightness and IRS).[22] As asset demand in the CM by future C-types determines the issue price, the resulting increase in liquidity is so strong that it makes the price of asset A upward sloping in its supply – at least, until that supply is so large that the force of diminishing marginal utility takes over.[23] But we are not done. When the supply of A becomes even larger, all OTC trade becomes concentrated in the market for A and B ceases to be liquid at all. As this happens, the price of asset A jumps upward discontinuously;

[22] To be precise: with IRS and $\delta_A = \delta_B$, we observe $e_C < e_N$ in the interior if and only if $A < B$. The more plentiful asset is more liquid.

[23] Weill (2008) has a result of similar flavor: he studies an extension of Duffie et al. (2005) with multiple assets, keeping the aggregate supply of tradable assets constant but allowing some assets to be in larger supply than others. He finds that the more plentiful assets are easier to find and have a higher price.

later, we will see that this effect of increasing returns provides a powerful incentive to the issuer of an asset to issue up to the point where competing assets are driven out of secondary markets.

There are three empirically relevant aspects of this theoretical result. First and most obviously, the upward sloping demand finds its counterpart in the observation that bond liquidity can be positively related to bond supply, which is well-established in the empirical literature on corporate bonds (Hotchkiss and Jostova, 2007; Alexander et al., 2000).

Second, this result is Step 2 of our explanation why two assets with otherwise similar features can have big differences in their liquidity. Even with a modest degree of IRS in matching, an asset in smaller supply is likely to be significantly less liquid than one which is in larger supply, as agents prefer to enter the market where gains from trade are larger, and through their own entry help to make this market "thick". And consider how this would interact with the first step described above: even with a small exogenous difference in market efficiency, the disadvantaged market is likely to see significantly less entry, and thereby becomes very "thin" indeed.[24] In the next section, we will see how these factors reinforce one another, and how they interact with an endogenous choice of supply.

Third, the result can help explain how consolidating secondary markets would be beneficial for asset issuers, a belief commonly held among practitioners (BlackRock, 2014). To see this, consider a version of our model with three issuers, A, B, and C. Compare the case where all bonds trade in distinct secondary markets with an alternative case where the OTC markets for bonds B and C merge. In our model, this would imply that, with even slight IRS in matching, the liquidity premium on bonds B and C will be higher in the second scenario, because the consolidation of the markets reduces the investors' risk of not being able to sell.

To summarize our results: we find that liquidity premia are always zero if asset supply is large but may be positive if asset supply is small enough. With CRS, the liquidity premium on a particular asset is always decreasing in that same asset's supply; but with IRS, liquidity depends positively on issue size and asset demand curves can therefore have *upward sloping* segments. However, the liquidity premium on an asset is always decreasing in the supply of *other* assets, which opens the door to strategic interaction.

[24] Interpreting market A as the market for U.S. Treasuries, there is an additional element that may add to this market's liquidity': the Federal Reserve (FED) often participates in this market by selling or buying large quantities of assets. For instance, in the period between November 2008 and September 2011, the FED purchased $\$1.19$ Trillion of Treasury debt, as part of a program now known as quantitative easing (QE). While our paper does not explicitly model interventions of the FED in the financial markets, in the form of open market operations or QE, it is reasonable to expect that the presence of a big player such as the FED in that market will be a pole of attraction for other investors, too: if I want to sell assets in the secondary market (like the C-types in our model) and I know that someone is purchasing billions worth of asset A in OTC_A, why would I go anywhere else? Readers who are interested in how one could model direct interventions of the FED in financial markets in a similar framework are referred to Herrenbrueck (2019a) and Geromichalos and Herrenbrueck (2017). For a careful empirical characterization of the effects of QE, see Song and Zhu (2018).

4 The economy with strategically chosen asset supply

4.1 The game between the asset issuers

We look at the non-cooperative game between two issuers who seek to maximize their utility. They live only in the CM, where they can work, consume, and issue assets. Their utility within the period is $\mathcal{Y}(X, H) = X - H$, where X, H denote consumption and work effort, and they discount the future by the same factor β as all agents. They take into account that the real price at which they can sell their asset, φp_j, depends on the supplies of both assets. For example, the problem of issuer A who has issued A^- assets in the previous period can be described by the following Bellman equation:

$$\begin{aligned} W^A(A^-) = \max_{X,H,A} \left\{ X - H + \beta W^A(A) \right\} \\ \text{s.t. } X + \varphi A^- = H + \varphi p_A A, \end{aligned}$$

which we can simplify to yield:

$$W^A(A^-) = -\varphi A^- + \max_{A} \left\{ \varphi p_A A + \beta W^A(A) \right\}.$$

Just like for private agents, the issuer's choice of A does not depend on their previous choices. We can use this, plus the fact that in steady state $\varphi/\hat{\varphi} = (1 + \mu)$, to solve for issuer A's objective:

$$\begin{aligned} J^A &= \frac{\varphi}{1+i} \left[(1+i) p_A - 1 \right] A \\ &= \frac{\varphi}{1+i} \big(\ell \, \alpha_{CA} \theta \left[u'(q_{1A}) - 1 \right] \big) A. \end{aligned} \tag{20}$$

With an analogous derivation, issuer B's objective is:

$$J^B = \frac{\varphi}{1+i} \big(\ell \, \alpha_{CB} \theta \left[u'(q_{1B}) - 1 \right] \big) B. \tag{21}$$

Simply put, each issuer seeks to maximize the product of the net liquidity premium L_j and the supply of their asset, taking into account that their choice of asset supply affects the general equilibrium choices of the agents.

The next step is to choose a solution concept for the game between the issuers. To keep our analysis simple and contained, we proceed in two stages that should be understood as distinct. First, we will describe the payoff structure facing the asset issuers in the stage game and analyze how this structure depends on parameters such as whether one of the assets has an exogenous liquidity advantage (δ_A, δ_B), and whether the matching function exhibits CRS or IRS (ρ). We believe this will allow our readers to extrapolate what kind of outcomes one might

obtain with their preferred solution concept.[25] Second, for the sake of concreteness, we will solve for static Nash equilibria of the stage game and analyze how these equilibria, as well as resulting macroeconomic outcomes and welfare, depend on the market parameters.

4.2 Strategic structure of the game

In this subsection, we analyze the strategic structure of the game (i.e., the incentives asset issuers face); in the next subsection, we will analyze how static Nash equilibria of the game depend on variations of this structure.

As before, we begin with the simplest case: "balanced CRS" in financial markets ($\rho = 0$ and $\delta_A = \delta_B$). As we saw in Proposition 1 above, the two corner equilibria are not robust to small errors: if a small measure of C-types happens to enter a market with no N-types, then N-types can profitably deviate by entering that market as well. More and more agents (C and N) will enter that market until the interior equilibrium is reached; consequently, in the case of balanced CRS, the interior equilibrium defined in part (e) of the Proposition is the interesting one to study. In this equilibrium, liquidity premia are positive, equal ($L_A = L_B > 0$), and depend only on the sum $A + B$: the assets are perfect substitutes and are priced along a common demand curve. This case of balanced CRS is therefore isomorphic to a version of the model where the assets could be traded in the *same* OTC market rather than in segmented markets as we assume here. And because the assets are perfect substitutes, the only Nash equilibrium of the game between the issuers is the symmetric Cournot equilibrium where both assets are issued in the same quantity, each approximately one-third of the quantity $\bar{D}$ that would drive the liquidity premium to zero.

Next, we are interested in the effects of exogenous liquidity differences. Specifically, we set δ_A equal to 1 and let δ_B vary, while maintaining CRS. Figure 5 illustrates the numerical results: the leftmost column shows the balanced CRS case, and the rightmost column confirms that if B has too much of a disadvantage, the interior equilibrium ceases to exist and all OTC trade is in the A-market. Issuer A gets to issue the monopoly quantity, approximately one-half of $\bar{D}$, and issuer B issues an arbitrary amount because asset B is illiquid in any case.

The intermediate values of δ_B, where the B-market is only a little bit worse than the A-market, show the transition. As we had already seen in Figure 4 (middle column), the demand curve for asset A has a kink whenever δ_B is less than δ_A. As long as δ_B is close enough, the Cournot-style interior equilibrium survives. When δ_B becomes too small, however, A prefers

[25] As one example, one may argue that political agents for whom liquidity rent is not the only consideration, such as the U.S. Treasury, are not Nash players but are able to precommit. We do explore this possibility in the Web Appendix: first, a "semi-strategic" case where the supply of A is set non-strategically, and issuer B best-responds to it; second, a Stackelberg duopoly where issuer A moves first and precommits to a (typically, large) issue size before B best-responds. And if we take the repeated interaction between the issuers seriously, there are even more possibilities, but they go beyond the scope of our paper.

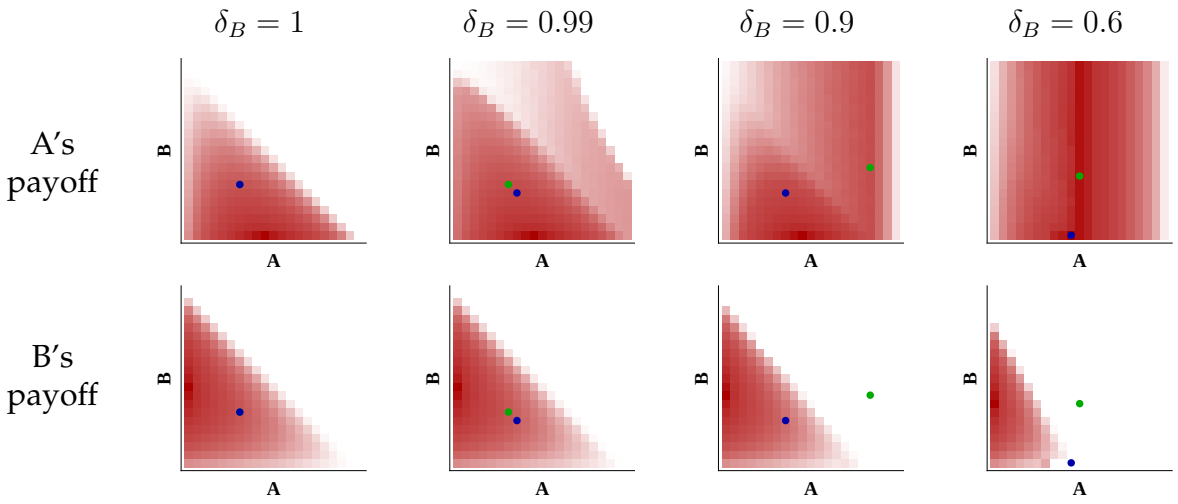


Figure 5: Payoffs as functions of asset supplies, with CRS ($\rho = 0$) and asset A having an exogenous liquidity advantage over asset B ($\delta_B \leq \delta_A = 1$). Darker shades of red indicate larger payoffs, white indicates zero. The blue and green dots indicate particular Nash equilibria.

to jump to a very large quantity that concentrates OTC trade in the A-market and drives B's liquidity premium to zero – even at the expense of a low liquidity premium for A itself. We will analyze the consequences for the economy in more detail later, but we can already see that the total supply of liquid assets is largest if B is somewhat illiquid, smallest if B is very illiquid, and in between if both A and B are very liquid.

To summarize: with CRS in financial markets, the structure of the game resembles Cournot competition. If not too unbalanced, CRS promotes the interior equilibrium in OTC markets where every asset is somewhat liquid. As a result, we see relatively small issue sizes.

Finally, we look at how the issuers' incentives are affected by IRS in financial markets. Specifically, we set $\delta_A = \delta_B = 1$ and let ρ vary. These results are illustrated in Figure 6; the leftmost column repeats the balanced CRS case from the previous figure, and the rightmost column illustrates how a strong degree of IRS makes the symmetric interior entry equilibrium so unstable that it is never reached as the subgame of the issuers' game. Why? Let us go back to Figure 3. Suppose that asset supplies are small and the interior entry equilibrium is played – i.e., both OTC markets are active. Issuer A has a strong incentive to supply more: yes, this moves her down her own demand curve (reducing her profits), but at the same time, the bigger surpluses in the A-market attract so many traders that the B-market shuts down (increasing A's profits). Of course, B has a symmetric incentive. With IRS, traders prefer to concentrate in one market, so the reward to issuers of offering a bigger trading surplus than their competitor becomes enormous.[26] Consequently, there is a (numerically approximate) Nash equilibrium where quantity A is so close to $\bar{D}$ that issuer B does not find it profitable to issue any more,

[26] Recall: when computing equilibria, we made the tie-breaking assumption that traders are more likely to pick the corner of the asset of which there is a larger supply. See the discussion at the end of Section 3.3.

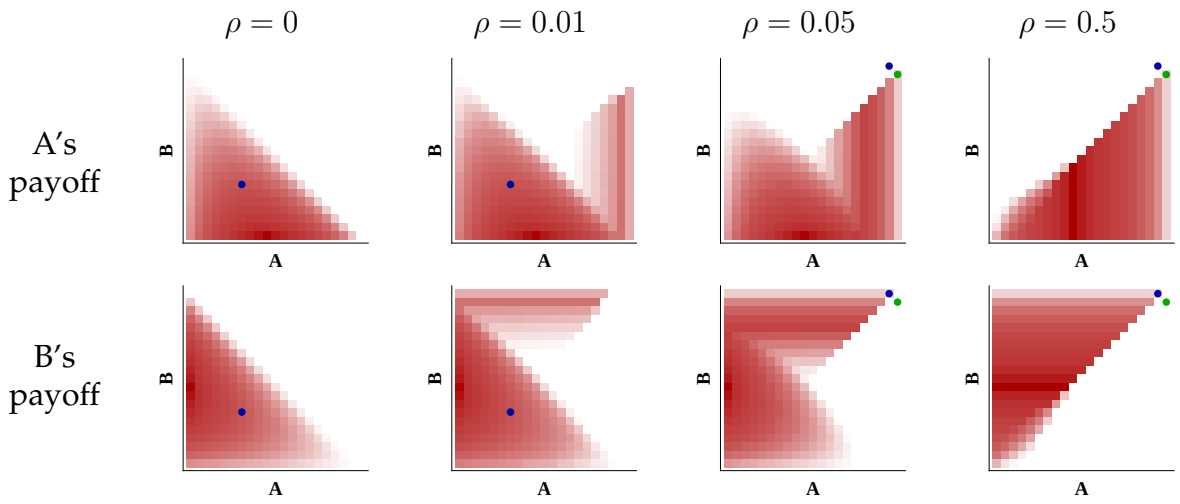


Figure 6: Payoffs as functions of asset supplies, with IRS but no exogenous liquidity advantage ($\delta_A = \delta_B = 1$). Darker shades of red indicate larger payoffs, white indicates zero. The green and blue dots indicate particular Nash equilibria.

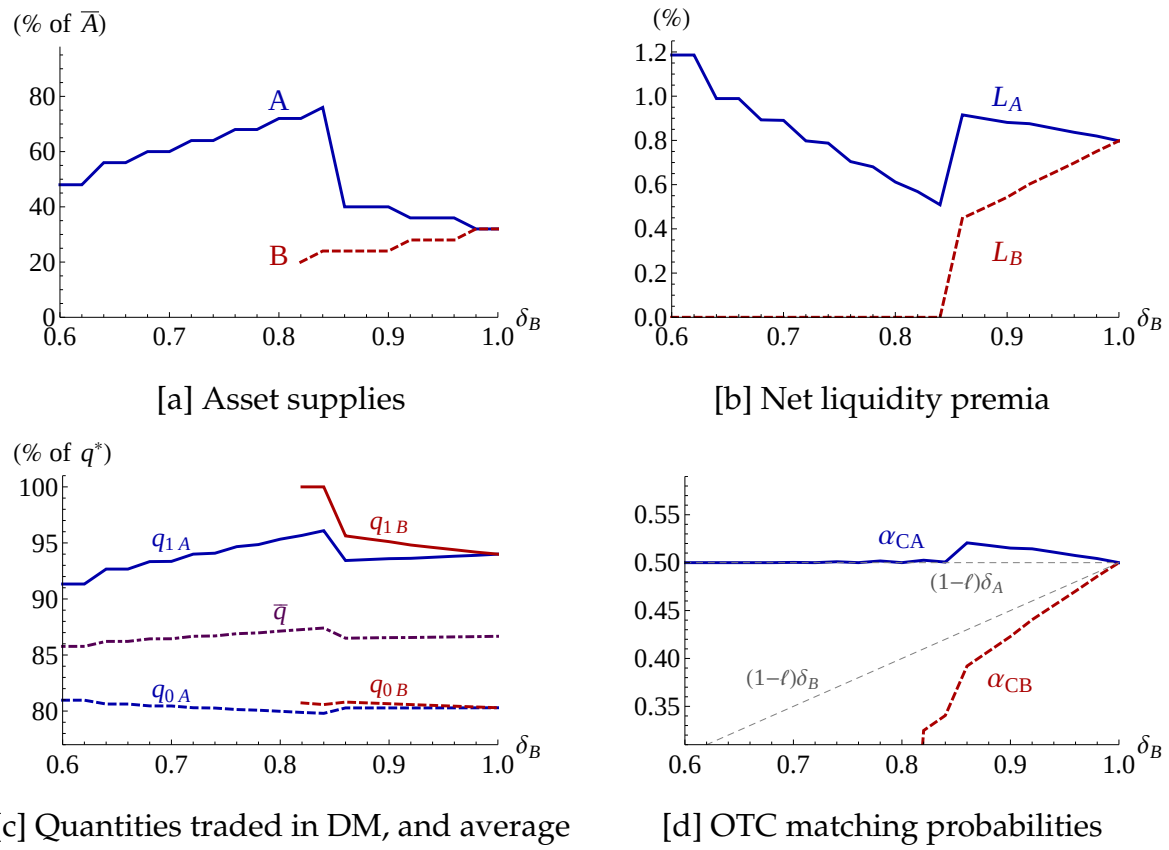
because in either case their asset would trade at a zero liquidity premium, either due to being illiquid or due to being plentiful. And there is a mirror Nash equilibrium with A and B's roles reversed, which is what the two dots in the top right corner of Figure 6 indicate.

For intermediate values of ρ, we see a smooth transformation of the playing field. For low ρ, assets tend to be strategic substitutes where issuers prefer to issuer neither too little nor too much, but for high ρ, assets become strategic complements where issuers strongly prefer to issue *more* than the other. Crucially, the fact that the playing field changes smoothly does not mean that the Nash equilibria change smoothly. On the contrary: as ρ increases, we get a jump transition from Cournot-type equilibria of low issue sizes and both assets being liquid to asymmetric equilibria of high issue sizes and only one asset being liquid. Note that the critical amount of IRS is approximately $\rho = 0.02$ – not particularly large – because it is the competition between issuers that gives a small amount of IRS a big endogenous 'kick'.

We can say that with enough IRS in financial markets, despite being a game in quantities rather than prices, the strategic structure of the game *resembles* Bertrand competition rather than Cournot. This promotes corner equilibria, where one asset ends up being very liquid and the other one not liquid at all. As long as there are no exogenous differences in market quality ($\delta_A \approx \delta_B$), in such equilibria the 'winning', liquid asset must be in large supply (close to $\bar{D}$).

4.3 Comparative statics

In this section, we analyze the comparative statics with respect to δ_B of the static Nash equilibria of the issuers' game. We consider both constant returns in matching and small amounts



[a] Asset supplies

[b] Net liquidity premia

[c] Quantities traded in DM, and average

[d] OTC matching probabilities

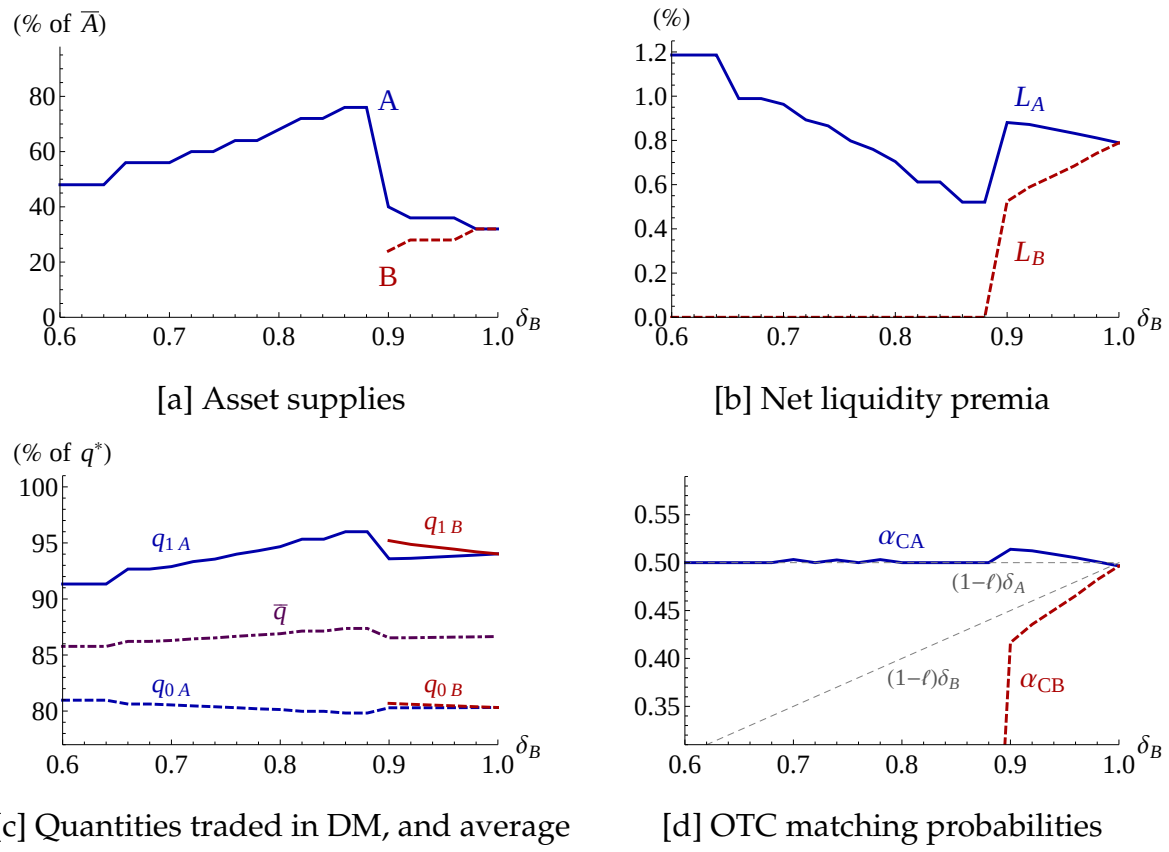
Figure 7: Comparative statics of the strategic equilibria with respect to δ_B, with CRS ($\rho = 0$).

of increasing returns. Our goal is to understand what happens if one of the assets (A for concreteness) has an exogenous matching advantage, and how the answer to this question interacts with the returns to scale in matching. Throughout this section, we hold $\delta_A = 1$ fixed.[27]

The case of CRS is illustrated in Figure 7. As δ_B declines slightly from 1 (the balanced case), A begins to issue more and B begins to issue less (panel [a]), but the strategic pattern of a Cournot game is maintained. The exogenous liquidity advantage of asset A is magnified by the entry choices of agents (panel [d]), which feeds back into a rising liquidity premium on asset A and a falling liquidity premium on asset B (panel [b]). Outputs diverge: C-types who hold asset A end up purchasing smaller quantities q_{0A} and q_{1A}, but the probability that they will obtain the larger one of the two, q_{1A}, increases. Conversely, C-types who still hold asset B despite its liquidity disadvantage are compensated with higher quantities q_{0B} and q_{1B} (panel [c]).

As δ_B declines further, we observe a discontinuity. At some point, the benefit to A from

[27] One detail to be aware of is how we compute the Nash equilibria. We iterate best responses of the two issuers on a finite grid of possible asset supplies which excludes asset supplies which we know can never give positive payoffs: zero and supplies exceeding $\bar{D}$. The starting point is the smallest positive asset supply on the grid (e.g., the point $(0.05\bar{D}, 0.05\bar{D})$ on a 20×20-grid). The remaining choice is whether we let A or B move first. In this section, all equilibria are computed with A moving first; the equilibria where B moves first are usually identical, payoff-identical, or mirror images. In Figures 5 and 6, Nash Equilibria where A moves first are indicated with a blue dot, and those where B moves first are indicated with a green dot.



[a] Asset supplies

[b] Net liquidity premia

[c] Quantities traded in DM, and average

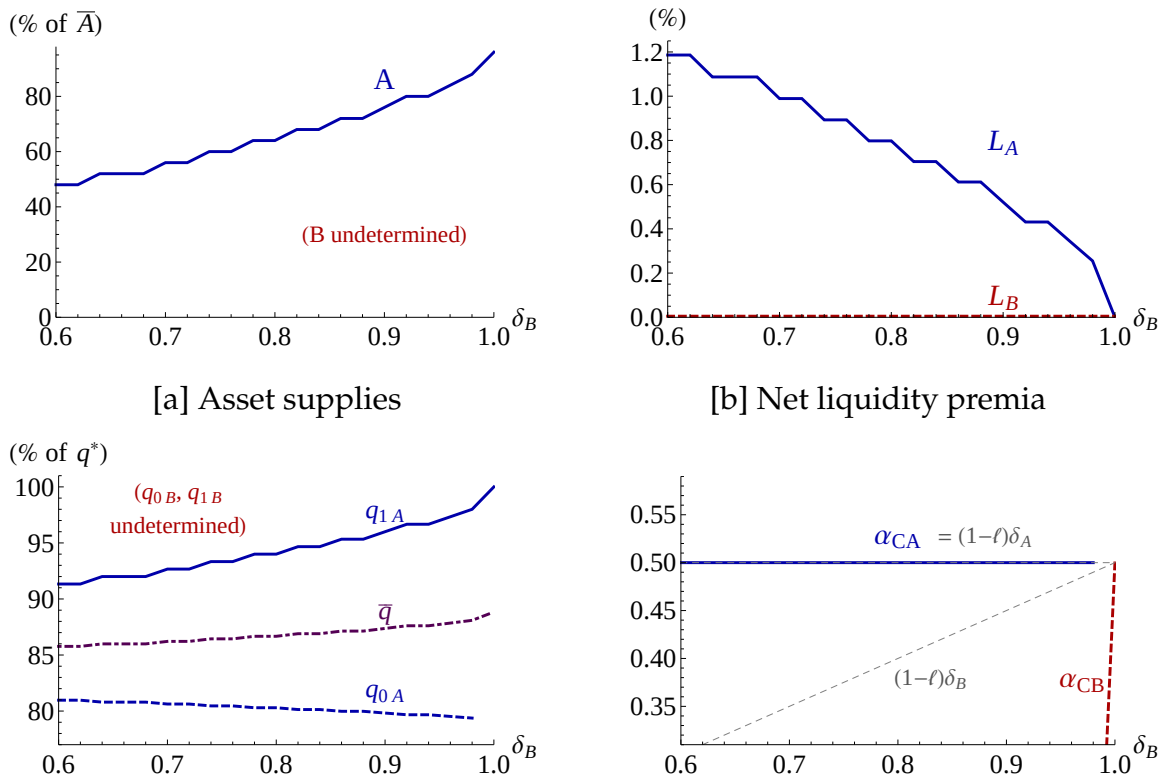
[d] OTC matching probabilities

Figure 8: Comparative statics with respect to δ_B, with just the tiniest bit of IRS ($\rho = 0.01$).

ramping up the issue size all the way to drive out B from the financial markets becomes too strong, so this is what A does. Asset B becomes fully illiquid, and therefore its issue size and the quantities q_{0B} and q_{1B} become indeterminate. As a result of this aggressive competition, average output of DM goods is highest at the discontinuity. If δ_B declines even more, the threat of trading asset B gradually diminishes; eventually, A becomes a monopolist who issues an intermediate quantity of asset A and average output declines to its lowest value. (Welfare is a more complicated story, as we explain in Section 4.4 below.)

When we allow for a very small degree of IRS in matching, $\rho = 0.01$ (illustrated in Figure 8), the results are almost identical to those with CRS, as one might expect given that ρ is so close to zero. Even so, we can see that the transition from the interior equilibrium to the A-corner where asset B is illiquid happens 'sooner', i.e., for a higher value of δ_B, than under CRS. Increasing returns make it slightly easier for A to drive B out of the market: in the example, A will do so for $\delta_B = 0.87$ under $\rho = 0.01$ but not under $\rho = 0$.

Based on Figure 6, one would guess that when increasing returns are strong enough, the Cournot-style equilibrium is eliminated in favor of aggressive competition for secondary market liquidity. But *how* strong do they need to be? Our perhaps surprising answer is: not very. As Figure 9 illustrates, the transition occurs somewhere between $\rho = 0.01$ and $\rho = 0.02$; in the



[a] Asset supplies

[b] Net liquidity premia

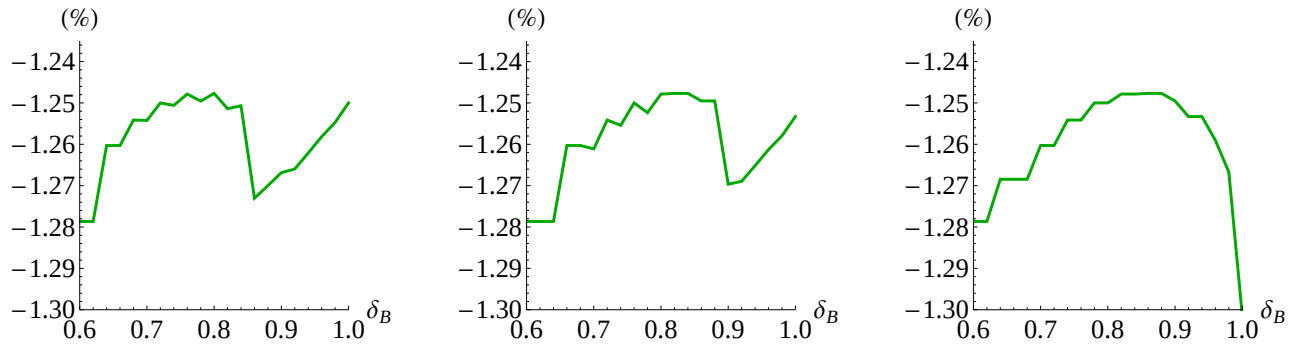
[c] Quantities traded in DM, and average

[d] OTC matching probabilities

Figure 9: Comparative statics with respect to δ_B, with a little bit more IRS ($\rho = 0.02$).

latter case, even with a relatively tiny degree of IRS, issuer competition is fierce and in every case only one market is open. However, this does not mean that the exogenous market quality parameters δ_A and δ_B stop mattering. When $\delta_B = \delta_A = 1$, an issuer who wishes to capture the secondary market must issue the quantity $\bar{D}$, which also drives her own payoff to zero. But as δ_B declines, so does the threat of B's competition, and therefore A's issuance is *negatively* related to her strategic advantage δ_A/δ_B.

While our model abstracts from a number of factors that are certainly influencing the borrowing decisions of the real-world issuers (the U.S. Treasury, large corporations, etc.), our theory generates solutions that resemble patterns in real-world asset markets. For instance, Figures 7 and 8 illustrate how even a small disadvantage of market B manifests itself as a higher matching probability for sellers of asset A (panel [d]), hence a larger liquidity premium for asset A (panel [b]), and how this mechanism is reinforced by issuer B's decision to scale back their issue size (panel [a]). The question whether the Treasury should be considered a strategic agent is interesting but not dispositive. In the Web Appendix, we consider a "semi-strategic" model where issuer B is strategic but issuer A is not. We show that the implications – at least, as far as issuer B is concerned – are broadly the same.



[a] $\rho = 0$ [b] $\rho = 0.01$ [c] $\rho = 0.02$

Figure 10: Welfare as a function of δ_B, measured as equivalent CM consumption, in percent deviations from the first-best.

4.4 The relationship between asset supplies, output, and welfare

We define social welfare $\mathcal{W}$ to be the total surplus across all DM trades, as follows:

$$\begin{aligned}\mathcal{W} \equiv \ell e_C \Big((1-\alpha_{CA}) \left[u(q_{0A}) - q_{0A} \right] + \alpha_{CA}(1-\theta) \left[u(\tilde{q}_{1A}) - \tilde{q}_{1A} \right] + \alpha_{CA}\theta \left[u(q_{1A}) - q_{1A} \right] \Big) \\ + \ell(1-e_C) \Big((1-\alpha_{CB}) \left[u(q_{0B}) - q_{0B} \right] + \alpha_{CB}(1-\theta) \left[u(\tilde{q}_{1B}) - \tilde{q}_{1B} \right] + \alpha_{CB}\theta \left[u(q_{1B}) - q_{1B} \right] \Big)\end{aligned} \tag{22}$$

In monetary models of this kind, there is no general relationship between the supply of liquid assets and output or welfare. For example, consider the corner equilibrium where only the OTC market for asset A is open (or assume for a moment that A is the only asset). Applying a recent result by Herrenbrueck and Geromichalos (2017) and Huber and Kim (2017), it can be shown that welfare is a *decreasing* function of the asset supply in a neighborhood $(\bar{A} - \epsilon, \bar{A})$. Why? First, note that as A increases (but is still below $\bar{A}$), q_{0A} falls and $(q_{1A}, \tilde{q}_{1A})$ rise, so the effect on average output is ambiguous and depends on parameters. However, the welfare impacts of these changes are weighted by the marginal utility term $u'(q) - 1$. If A is close to $\bar{A}$, then $u'(q_{1A})$ is close to $u'(q^*) = 1$; thus, the welfare gain which successful traders receive from higher A vanishes, but the welfare loss of unsuccessful traders does not, and the overall welfare effect is negative. This is confirmed by combining panels [a] and [c] of Figure 9, showing increasing asset supply and output near $\bar{A}$, with panel [c] of Figure 10, which shows the drop in welfare.

Huber and Kim (2017) also show (in a model with a single OTC market and a single bond) that when both sides in the OTC market have some bargaining power, then welfare is an increasing function of the asset supply in a neighborhood around zero, so the optimal asset supply is nonzero. We confirm their result for a special case of our model in the next section, where the two OTC markets are identical, assets are therefore perfect substitutes, and we can solve the model analytically and in closed form.

What does this imply for the relationship between market microstructure and welfare in

general? First, using the fact that an asset supply close to $\bar{D}$ is always 'too much' from a welfare perspective, we argue that any condition that leads to aggressive competition among the asset issuers is best avoided. In particular, the general intuition that the more competition, the better for social welfare, is not valid when it comes to liquid assets. The same reasoning would apply when matching is CRS and we compare a Cournot oligopoly of few versus many competitors.

Second, there is less clarity when we are far from the aggressive "everyone issues $\bar{D}$" case. For intermediate bargaining power $\theta \approx 0.5$, a Cournot duopoly is better for welfare than a monopoly, so it is also possible to have too little competition. But the exact turning point will depend on details.

Third, and perhaps surprisingly, the effect of the exogenous 'market quality' parameter δ_B on welfare is not monotonic. In fact, for IRS and $\delta_B \approx \delta_A$, the effect is negative, because similarity promotes aggressive competition. For CRS, we have shown that $\delta_B \ll \delta_A$ promotes a monopoly and $\delta_B \approx \delta_A$ promotes a duopoly, but it is *intermediate* values of δ_B that promote the most aggressive competition, the largest supply of liquid assets, and a dip in welfare. It is also important to recognize that little of the welfare results can be ascribed to the direct effect on the extensive margin of OTC trade, as the [d]-panels of Figures 7-9 show: asset B is endogenously illiquid for $\delta_B < 0.9$, no OTC trade in that market actually takes place, but the threat that it might still affects the equilibrium.

5 The special case of balanced CRS

As we showed in part (e) of Proposition 1, in the special case of $\rho = 0$ and $\delta_A = \delta_B \equiv \delta$ there is an equilibrium of the economy with $e_{_C} = e_{_N} = A/(A+B)$, and symmetry in the other equilibrium variables: $q_{0A} = q_{0B}$, $q_{1A} = q_{1B}$, and $p_A = p_B$. So we can drop the asset subscripts for the rest of this section. If we further assume that $u(q) \equiv \log(q)$, which normalizes the first-best level of DM production to $q^* = 1$, there is a closed-form solution both for the portfolio-choice subgame and for the Cournot-Nash equilibrium of the issuers. In this section, we analyze this closed-form solution in more detail as it provides valuable intuition for the more complex, asymmetric, cases analyzed in the previous section.

First, define the parameter $\kappa \equiv (1-\ell)\delta\theta$, which summarizes financial market liquidity from a C-type's point of view. (The $(1-\ell)$-term is the measure of N-types in the economy, and it enters here through the CRS matching function.) The upper bound on the overall asset supply where assets become abundant in OTC trade is $\bar{D} \equiv i/[\ell(1-\kappa)] \cdot M$, and for fixed asset supplies which satisfy $A + B < \bar{D}$, we obtain the solution:

$$q_0 = \frac{1-\kappa+\kappa\frac{M}{M+A+B}}{1+i/\ell}, \quad q_1 = \frac{1+(1-\kappa)\frac{A+B}{M}}{1+i/\ell}, \quad p = \frac{1}{1+i}\cdot\left(1+\ell\kappa\,\frac{i/\ell-(1-\kappa)\frac{A+B}{M}}{(1-\kappa)\frac{M+A+B}{M}+\kappa}\right)$$

For large enough asset supplies, i.e., $A + B \geq \bar{D}$, it is easy to solve for $q_0 = [1 + \bar{D}]^{-1}$, $q_1 = 1$, and $p = 1/(1 + i)$.

From here on, we could move directly to the solutions to the issuer's game; they will be in closed form, but nonlinear. We can obtain something even simpler by recognizing that while the asset price p is not exactly linear in the asset supplies, it is nearly so as long as the asset supplies are small enough. To formalize this idea, we note that as $i \searrow 0$, we have $(q_0, q_1) \nearrow q^*$ and also $\bar{D} \searrow 0$. Thus, we define counterparts to the above equilibrium objects which are log-linearized around the first-best point, and indicate them with a circumflex:[28]

$$\hat{q}_0 \equiv i \cdot \frac{d}{di} \log(q_0/q^*), \qquad \hat{q}_1 \equiv i \cdot \frac{d}{di} \log(q_1/q^*), \qquad \hat{p} \equiv i \cdot \frac{d}{di} \log(p)$$

For the convenience of shorter formulas, we define the supply of assets relative to money to be $s \equiv (A + B)/M$ (with the obvious extension to many issuers: $s \equiv \sum_j A_j/M$), and define its upper bound where the liquidity premium becomes zero to be: $\bar{s} \equiv \bar{D}/M = i/[\ell(1 - \kappa)]$. For the interior case where $s \leq \bar{s}$, we obtain:

$$\hat{q}_0 = -\frac{i}{\ell} - \kappa s, \qquad \hat{p} = L - i,$$
$$\hat{q}_1 = -\frac{i}{\ell} + (1 - \kappa)s, \qquad \text{where: } L = \kappa i - \ell(1 - \kappa)\kappa s$$

Now, the effective demand curve for liquid assets $L(s)$ is linear in asset supplies. Solving for monopoly, Cournot duopoly, Stackelberg duopoly, and competitive (price-taking) issue sizes is thus straightforward.[29] Writing asset supplies as proportions of the upper bound $\bar{s}$:

	Total asset supply s	Liquidity premium L
Monopoly	$\frac{1}{2}\bar{s}$	$\frac{1}{2}\kappa i$
Cournot	$\frac{2}{3}\bar{s}$	$\frac{1}{3}\kappa i$
Stackelberg	$\frac{3}{4}\bar{s}$	$\frac{1}{4}\kappa i$
Competition	$\bar{s}$	0

Quantities of goods $\hat{q}_0$ and $\hat{q}_1$ corresponding to each of the above asset supply solutions can be computed easily. Welfare is a little bit more interesting. Recall the definition of welfare as

[28] What about the quantity $\tilde{q}_1$, which is the amount of DM-production that the C-type can afford after the N-type makes the offer in the OTC market? It turns out that near the Friedman rule, $\tilde{q}_1 \approx q_1$ and they have the same Taylor expansion around q^*. So $\hat{q}_1$ is the approximate post-trade quantity no matter whether the C-type or the N-type made the offer.

[29] Crucially, all of these solutions assume that issuers have zero marginal cost of issuing bonds. This may not be realistic; perhaps what distinguishes large borrowers such as the Treasury from smaller competitors is precisely that the Treasury has a lower cost of creating safe assets (which then have the chance to become liquid by virtue of being traded in thick secondary markets). In the context of our general model with IRS and/or unequal markets, we explore this extension numerically in the Web Appendix.

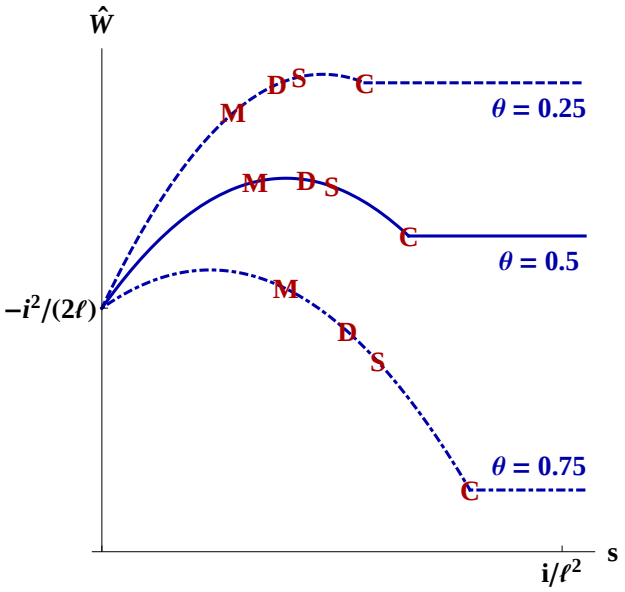


Figure 11: Approximate social welfare in the balanced CRS case ($\rho = 0$, $\delta_A = \delta_B = 1$) and how it depends on bargaining power θ and scaled asset supply $s = (A + B)/M$. Possible outcomes of the issuers' game are indicated: (M)onopoly, Cournot (D)uopoly, (S)tackelberg duopoly, and price-taking (C)ompetition.

total surplus across all trades, $\mathcal{W}$, in Equation (22). Since $\mathcal{W}$ as a function of i is flat at the Friedman rule, we need to take a second-order Taylor expansion. Net of the first-best welfare amount (which equals $-\ell$), we obtain the approximation:

$$\hat{\mathcal{W}} \equiv -\frac{\ell}{2}\Big([1-\delta(1-\ell)]\,\hat{q}_0^2 + \delta(1-\ell)\,\hat{q}_1^2\Big)$$

We illustrate this welfare function in Figure 11. Its shape is critically influenced by θ, the bargaining power of the C-type in OTC trade. A higher value of θ (through κ) directly decreases both quantities $(\hat{q}_0, \hat{q}_1)$, because if liquidity is easier to obtain ex-post, agents will obtain less of it ex-ante. Thus, the best value of θ for social welfare is the lowest one: $\theta = 0$. But taking θ as given, its level also governs how much asset supply is best for welfare. Specifically, call this level s^*; after some algebra, it turns out that it is:

$$s^* = \frac{(1-\theta)[1-\theta\delta(1-\ell)]}{1-(2-\theta)\theta\delta(1-\ell)} \cdot \bar{s},$$

where the ratio term is bounded between zero and one. As a concrete example, for all numerical experiments in this paper we used the parameters $\theta = \ell = 0.5$; thus, depending on market quality δ, the second-best total asset supply is between 50% and 60% of its upper bound (where the liquidity premium becomes zero). This is just above the monopoly outcome and just below the Cournot duopoly outcome, which explains why the best welfare outcomes in Figure 10 are obtained for a modest degree of competition.

6 Conclusion

We develop a model in which an asset's liquidity and, hence, its equilibrium price depend on:

1. The microstructure of the secondary market where that asset trades;
2. The microstructure of the secondary market(s) where "competing" assets trade;
3. The decision of agents to visit these secondary markets; (which in turn depends on the microstructure of the various markets), and;
4. The endogenous supply of the various assets.

Our model delivers a number of new insights. Even with small amounts of increasing returns, asset demand curves can be upward sloping because IRS encourages market concentration and agents are more likely to concentrate in market of an asset with plentiful supply. We also show that small differences in the microstructure of an OTC market can be magnified into a big endogenous liquidity advantage for one asset, because traders would prefer to be in the thick market, and through their own entry help make it even thicker.

Our model predicts that for a reasonable set of parameters, a big and well-established borrower such as the Treasury can enjoy a significant liquidity advantage, to the point where they may be the *only* issuer of assets that trade at a liquidity premium. But in our model, whether the Treasury will be a monopolist in the issuance of liquid assets or not is endogenous. The model describes the conditions under which this will be the case, but it also describes what it would take for the Treasury to lose this liquidity advantage. A more efficient and consolidated secondary market for corporate (or municipal) bonds – as recently advocated by market analysts (BlackRock, 2014) – is one such candidate: it would increase the secondary market liquidity for corporate bonds and, thus, jeopardize the monopoly of the Treasury.

Appendix

A.1 Discussion of key modeling choices

One of the motivating examples of this paper is the superior liquidity of Treasuries over equally safe debt of similar characteristics. One question that arises is whether other forms of debt can be as safe as Treasuries. The short answer is yes: AAA corporate or municipal bonds are considered equally safe. AAA-rated bonds have an exceptional degree of creditworthiness, because the issuer can easily meet its financial commitments. In many cases, AAA corporate bonds are secured by assets (e.g., equipment, machinery or real estate) that are pledged as collateral, and the creditor has a claim on the collateral if the issuer defaults on the bond. Similarly, in the case of the so-called "insured municipal bonds", which are also AAA-rated, the investor is covered if the issuer fails to make coupon or principal payments. Finally, a large number of empirical papers (Krishnamurthy and Vissing-Jorgensen (2012) would be an excellent example) estimate the liquidity premium of Treasuries over high quality corporate bonds (with similar characteristics) by focusing only on bonds that are considered *as safe as Treasuries*.[30]

We adopt a matching function that admits both CRS and IRS as subcases, and we present results for each case, but one may say that some of the most interesting results of the paper are derived using IRS. So how realistic are increasing returns to scale in financial markets? Quite realistic, in fact, which is well-established both at the theoretical and the empirical level. Duffie et al. (2005), and the vast majority of papers that follow their seminal work, adopt an IRS matching technology.[31] Furthermore, a number of empirical finance papers seem to confirm the relevance of IRS in OTC markets: for example, there is strong evidence that markets with higher trading volumes have lower bid-ask spreads. (See the discussion on page 54 of Vayanos and Wang (2012).) If higher spreads are associated with longer search times, this would be an argument suggesting IRS, since it would imply that markets with higher trading volume have more traders who are searching, and have lower bid ask spreads because trading delays are shorter. Are higher bid-ask spreads indeed associated with longer search times? Any theoretical model of OTC trade we are aware of would predict so.[32] As for the data, Amihud and Mendelson (1986) provide support in favor of this empirical regularity. Also, a quick glance

[30] Another example of assets that are equally safe but have different degrees of liquidity is (common) Treasuries and Treasury inflation protected securities (TIPS). It is clear that the default risk of these two types of assets is identical (they are issued by the same authority). Nevertheless, a large empirical finance literature documents that TIPS suffer a significant illiquidity discount compared to common Treasuries (after controlling for expected inflation). See for example Andreasen, Christensen, Cook, and Riddell (2016).

[31] In that paper, the total number of matches between buyers and sellers of assets is given by $2\lambda\mu_B\mu_S$, where μ_B, μ_S are the respective measures of buyers and sellers, equivalent to $\rho = 1$ in our model. Hence, the arrival rate of a buyer to a seller is $2\lambda\mu_B$, which does not depend of the number of sellers. This process is therefore not just IRS, but completely congestion-free.

[32] For example, in Duffie et al. (2005), the bid-ask spread is strictly decreasing in the arrival rate of trading opportunities. Faster arrival rates imply a better outside option for the investor, thus a better bargaining position.

at some of the main OTC markets suggests that this relationship is indeed true: markets that are characterized by long trading delays, e.g., municipal bonds, are also typically characterized by large spreads. Crucially, the amount of IRS needed in order to explain big divergences in market outcomes in our model is really small.

A key assumption in our analysis is that secondary markets are segmented and agents can visit only one per period. The first part of this assumption is certainly realistic: Treasuries and municipal (or corporate) bonds do trade in secondary markets that are completely distinct. The second part of the assumption, according to which agents can visit only *one* market, is stronger – but, clearly, it is not meant to be taken literally. It does not imply a significant loss of generality since it is a qualitative rather than quantitative ingredient for our model's central mechanism: it is just a stark way to capture the idea that even if some investors do visit multiple markets, they will visit the market where they expect to find better trading conditions more frequently.[33, 34]

In the absence of market segmentation, the assets would be perfect substitutes and their prices would always be identical. But the empirical finance literature abounds with examples of assets that have pretty much identical characteristics, yet they trade at significantly different prices (and in secondary markets with very different levels of liquidity, as measured by bid-ask spreads, trading volumes, etc). It is also a fact that most fixed income dealers simply do not intermediate multiple kinds of securities. Clearly there is some cost to becoming an expert in a specific security, even for such similar ones as Treasury and AAA corporate bonds. Thus, market segmentation is not only essential for our results, but also the empirically relevant case.

In our model, we study the differentiated Cournot game played by two bond issuers. One question that arises is whether in reality bond issuers are strategic. ('Strategic' has two relevant meanings: whether the issuers' objective includes profit/rent maximization, and whether they have market power.) First, the quote of the Assistant Secretary of the Treasury (presented in Footnote 2) clearly indicates that the Treasury is interested in maximizing its rent from debt issuance (although they call it "minimizing borrowing costs"). Similar evidence can be found for debt issuing corporations. Greenwood, Hanson, and Stein (2010) document that debt issuing corporations pay close attention to the actions taken by the Treasury and respond to these moves by filling in the supply gaps created by changes in government financing patterns. For another example, Robert Tipp, the Managing Director and Chief Investment Strategist of Prudential Investment Management, highlights that chief financial officers in big corporations are

[33] In reality, every time an investor buys or sells assets she has to incur some fixed cost. (This may include the cost of acquiring information about the asset to be traded, the time and effort spent to locate a trading partner, or commission fees.) If such a fixed cost was introduced into the model, an agent who wishes to boost her liquidity would typically try to avoid visiting two markets, and she would only visit the market with better conditions. Hence, the more complex model could deliver as a result what our simpler model adopts as an assumption: that each agent liquidates assets only in one OTC market.

[34] This assumption is analogous to a discrete choice model; such models are popular in economics (especially in Industrial Organization) because they offer a tractable way of modeling individual consumer choice over various goods while permitting a realistic description of aggregate market shares.

paying close attention to the market conditions and especially to the demand for bonds issued by the biggest player in the market: the Treasury.[35]

Given the discussion so far, we think that writing down a model where issuers play a differentiated Cournot game is a reasonable choice. In the baseline model, we focus on the case of a duopoly, but one can always generalize the model to include a general number of issuers, $N > 2$. Then, how much market power each issuer has is a (decreasing) function of N. Overall, we think that a model where issuers play a Cournot game is closer to reality than a model where issuers behave competitively.[36] Crucially, our baseline model can be easily extended in order to study a variety of alternative market structures. The Web Appendix contains three (out of many possible) such extensions: B.1 studies equilibrium in the presence of a non-strategic issuer with fixed supply; B.2 studies Stackelberg equilibria where issuer A is the first mover; B.3 allows for issuer B to have a positive marginal cost of issuing assets.

A.2 Value functions, and terms of trade in the frictional markets

A.2.1 Value functions

We begin with the description of the value functions in the CM. Consider first a buyer who enters this market with m units of fiat money and d_j units of asset $j = \{A, B\}$. The Bellman equation of the buyer is given by:

$$W(m, d_A, d_B) = \max_{\substack{X,H,\hat{m},\\ \hat{d}_A,\hat{d}_B}} \left\{ X - H + \beta\, \mathbb{E}_i \left\{ \max \left\{ \Omega^i_A\left(\hat{m}, \hat{d}_A, \hat{d}_B\right), \Omega^i_B\left(\hat{m}, \hat{d}_A, \hat{d}_B\right) \right\} \right\} \right\}$$
$$\text{s.t. } X + \varphi(\hat{m} + p_A\hat{d}_A + p_B\hat{d}_B) = H + \varphi(m + \mu M + d_A + d_B),$$

where variables with hats denote portfolio choices for the next period, and $\mathbb{E}$ denotes the expectations operator. The price of money is expressed in terms of the general good but the price of bonds is expressed in nominal terms. The function Ω^i_j represents the value function in the OTC market for asset $j \in \{A, B\}$ for a buyer of type $i \in \{C, N\}$, to be described in more detail below. At the optimum, X and H are indeterminate but their difference is not. Using this fact and substituting $X - H$ from the budget constraint into W yields:

$$W(m, d_A, d_B) = \varphi(m + \mu M + d_A + d_B) \quad \ldots$$

[35] Source: http://www.marketwatch.com/story/treasury-yields-edge-higher-apple-expected-to-issue-bonds-2016-02-16.

[36] For instance, in September 2013, Verizon issued bonds worth 49 billion dollars; in January 2016, Anheuser-Busch InBev issued bonds worth 46 billion dollars; in March 2018, CVS issued bonds worth 40 billion dollars (the list goes on). It would be hard to argue that when these corporations issue debt of this size they behave as measure zero agents whose actions have no effect on market prices.

$$+ \max_{\hat{m},\hat{d}_A,\hat{d}_B} \Big\{ -\varphi(\hat{m} + p_A\hat{d}_A + p_B\hat{d}_B) + \beta\ell \max\left\{\Omega_A^C\left(\hat{m},\hat{d}_A,\hat{d}_B\right), \Omega_B^C\left(\hat{m},\hat{d}_A,\hat{d}_B\right)\right\} + \beta(1-\ell)\max\left\{\Omega_A^N\left(\hat{m},\hat{d}_A,\hat{d}_B\right), \Omega_B^N\left(\hat{m},\hat{d}_A,\hat{d}_B\right)\right\}\Big\}. \tag{A.1}$$

In the last expression, we have also used the fact that the representative buyer will be a C-type with probability ℓ in order to replace the expectations operator. As is standard in models that build on LW, the optimal choice of the agent does not depend on the current state (due to the quasi-linearity of $\mathcal{U}$), and the CM value function is linear. We write:

$$W(m, d_A, d_B) = \varphi(m + d_A + d_B) + \Upsilon, \tag{A.2}$$

where the constant Υ collects the remaining terms that do not depend on the state variables m, d_A, d_B.

As is well-known, a seller will not wish to leave the CM with positive amounts of money and bond holdings. Therefore, when entering the CM a seller will only hold money that she received as payment in the preceding DM, and her CM value function is given by:

$$W^S(m) = \max_{X,H}\left\{X - H + V^S\right\}$$
$$\text{s.t. } X = H + \varphi m,$$

where V^S denotes the seller's value function in the forthcoming DM. We can again use the budget constraint to substitute $X - H$ and show that W^S will be linear:

$$W^S(m) = \varphi m + V^S \equiv \Upsilon^S + \varphi m. \tag{A.3}$$

We now turn to the description of the OTC value functions. Recall that $e_{_C} \in [0,1]$ and $e_{_N} \in [0,1]$ denote the fraction of C-types and N-types, respectively, who are entering OTC$_A$. Using the matching probabilities α_{ij} defined in Equations (1)-(2), we can now define the value function for an agent of type $i = \{C, N\}$ who decides to enter OTC$_j$, $j = \{A, B\}$. Let ζ_j, denote the amount of money that gets transferred to the C-type, and χ_j the amount of assets (of type j) that gets transferred to the N-type in a typical match in OTC$_j$, $j = \{A, B\}$. These terms are described in detail in Lemma A.2 below. We have:

$$\Omega_A^C(m, d_A, d_B) = \alpha_{CA} V\left(m + \zeta_A, d_A - \chi_A, d_B\right) + \left(1 - \alpha_{CA}\right) V(m, d_A, d_B), \tag{A.4}$$
$$\Omega_B^C(m, d_A, d_B) = \alpha_{CB} V\left(m + \zeta_B, d_A, d_B - \chi_B\right) + \left(1 - \alpha_{CB}\right) V(m, d_A, d_B), \tag{A.5}$$
$$\Omega_A^N(m, d_A, d_B) = \alpha_{NA} W\left(m - \zeta_A, d_A + \chi_A, d_B\right) + \left(1 - \alpha_{NA}\right) W(m, d_A, d_B), \tag{A.6}$$
$$\Omega_B^N(m, d_A, d_B) = \alpha_{NB} W\left(m - \zeta_B, d_A, d_B + \chi_B\right) + \left(1 - \alpha_{NB}\right) W(m, d_A, d_B), \tag{A.7}$$

where V denotes a buyer's value function in the DM. Notice that N-type buyers proceed directly to next period's CM.

Lastly, consider the value functions in the DM. Let q denote the quantity of goods traded, and τ the total payment in units of fiat money. These terms are described in detail in Lemma A.1 below. The DM value function for a buyer who enters that market with portfolio (m, d_A, d_B) is given by:

$$V(m, d_A, d_B) = u(q) + W(m - \tau, d_A, d_B), \tag{A.8}$$

and the DM value function for a seller (who enters with no money or assets) is given by:

$$V^S = -q + \beta W^S(\tau).$$

A.2.2 The terms of trade in the OTC markets and the DM

Consider a meeting between a C-type buyer with portfolio (m, d_A, d_B) and a seller who, in the beginning of the DM sub-period, holds no money or assets. The two parties bargain over a quantity q to be produced by the seller and a cash payment τ, to be made by the buyer. The buyer makes a TIOLI offer maximizing her surplus subject to the seller's participation constraint and the cash constraint. The bargaining problem can be described by:

$$\begin{aligned} \max_{\tau,q} \{u(q) + W(m - \tau, d_A, d_B) - W(m, d_A, d_B)\} \\ \text{s.t. } -q + W^S(\tau) - W^S(0) = 0, \end{aligned}$$

and the cash constraint $\tau \leq m$. Substituting the value functions W, W^S from (A.2) and (A.3) into the expressions above, allows us to simplify this problem to:

$$\begin{aligned} \max_{\tau,q} \{u(q) - \varphi\tau\} \\ \text{s.t. } q = \varphi\tau, \end{aligned}$$

and $\tau \leq m$. The solution to the bargaining problem is described in the following lemma.

Lemma A.1. *Let m^* denote the amount of money that, given the CM value of money, φ, allows the buyer to purchase the first-best quantity q^*, i.e., let $m^* = q^*/\varphi$. Then, the solution to the bargaining problem is given by $\tau(m) = \min\{m, m^*\}$ and $q(m) = \varphi \min\{m, m^*\}$.*

Proof. The proof is standard and it is, therefore, omitted. □

The solution to the bargaining problem is straightforward. The only variable that affects the solution is the buyer's money holdings. As long as the buyer carries m^* or more, the first-

best quantity q^* will always be produced. If, on the other hand, $m < m^*$, the buyer does not have enough cash to induce the seller to produce q^*. The cash constrained buyer will give up all her money, $\tau(m) = m$, and the seller will produce the quantity of good that satisfies her participation constraint under $\tau(m) = m$, namely, $q = \varphi m$.

While Lemma A.1 describes the bargaining solution for all possible money holdings by the C-type buyer, we know that, since $\mu > \beta - 1$, the cost of carrying money is strictly positive and a buyer will never choose to hold $m > m^*$.[37] Hence, from now on we will focus on the binding branch of the bargaining solution, i.e., we will set $\tau(m) = m$ and $q(m) = \varphi m$.

We now describe the terms of trade in the OTC markets. Consider a meeting in OTC$_j$, $j = \{A, B\}$, between a C-type carrying the portfolio (m, d_A, d_B) and an N-type with portfolio $(\tilde{m}, \tilde{d}_A, \tilde{d}_B)$. These agents negotiate over an amount of money, ζ_j, to be transferred to the C-type, and an amount of type-j assets, χ_j, to be transferred to the N-type. Recall that the C-type makes a TIOLI offer to the N-type with probability θ, and vice versa. In the match under consideration, the surpluses for the C-type and the N-type agents are given by:

$$\begin{aligned} S_{Cj} &= V(m + \zeta_j, d_A - \mathbb{I}\{j = A\}\,\chi_A, d_B - \mathbb{I}\{j = B\}\,\chi_B) - V(m, d_A, d_B) \\ &= u(\varphi(m + \zeta_j)) - u(\varphi m) - \varphi\chi_j, \qquad \text{(A.9)} \\ S_{Nj} &= W(\tilde{m} - \zeta_j, \tilde{d}_A + \mathbb{I}\{j = A\}\,\chi_A, \tilde{d}_B + \mathbb{I}\{j = B\}\,\chi_B) - W(\tilde{m}, \tilde{d}_A, \tilde{d}_B) = \varphi(\chi_j - \zeta_j), \qquad \text{(A.10)} \end{aligned}$$

where $\mathbb{I}$ denotes the identity function, and the second equalities in the equations above exploit the definitions of the functions V, W (i.e., Equations (A.8) and (A.2), respectively).

Consider first the case in which the C-type makes the TIOLI offer. Then, the bargaining problem is equivalent to maximizing S_{Cj} (with respect to ζ_j, χ_j), subject to $S_{Nj} = 0$ and $\chi_j \leq d_j$. On the other hand, if it is the N-type who makes the offer, the problem is equivalent to maximizing S_{Nj}, subject to $S_{Cj} = 0$ and $\chi_j \leq d_j$.

We restrict attention to equilibria where the N-type's money holdings never limit the trade, hence the corresponding constraint $\zeta_j \leq \tilde{m}$ is slack. A sufficient condition that guarantees this in equilibrium is given by inequality (6): inflation rates must be low enough that C-types (who carry m units of money) and N-types (who carry $\tilde{m}$) can always obtain the first-best m^* if they were to pool their money ($m + \tilde{m} \geq m^*$). Actual trade may achieve m^* or not, depending on whether the C-type carries enough assets to compensate the N-type for her money. Excluding the scarce-money branch of the bargaining solution is convenient: that branch ultimately generates a kink in the value function, which gives rise to an asset pricing indeterminacy, as we extensively analyzed in Geromichalos and Herrenbrueck (2016). It is also innocent for the purposes of our present paper: assets can only be priced (in the CM) at a determinate liquidity

[37] Even if the buyer in question matches with an N-type in the preceding OTC round and acquires some extra liquidity, she will never choose to adjust her post-OTC money balances in a way that these exceed m^*. This would be unnecessary since carrying m^* is already enough to buy her the first-best quantity in the forthcoming DM.

premium if $\chi_j \leq d_j$ binds (in the OTC) but $\zeta_j \leq \tilde{m}$ does not. Since our interest is in asset issuers who seek to exploit a positive premium, we think the restriction is acceptable.

The solution to the bargaining problem is described in the following lemma.

Lemma A.2. *a) Suppose that the C-type is making the TIOLI offer. Define* $\bar{d}^C \equiv m^* - m$. *Then, the bargaining solution is given by* $\chi_j(m, d_j) = \zeta_j(m, d_j) = \min\{d_j, \bar{d}^C\}$.

b) Suppose that the N-type is making the TIOLI offer. Define $\bar{d}^N \equiv [u(q^*) - u(\varphi m)]/\varphi$. *Then, the bargaining solution is given by* $\chi_j(m, d_j) = \min\{d_j, \bar{d}^N\}$ *and:*

$$\zeta_j(m, d_j) \quad = \quad \begin{cases} \tilde{\zeta}_j(m, d_j), & \text{if } d_j < \bar{d}^N, \\ m^* - m, & \text{if } d_j \geq \bar{d}^N, \end{cases}$$

where we have defined:

$$\tilde{\zeta}_j(m, d_j) \equiv \{\zeta : u(\varphi(m + \zeta)) - u(\varphi m) - \varphi d_j = 0\}.$$

Proof. It is straightforward to check that the suggested answer satisfies the necessary and sufficient conditions for maximization in each case. □

The OTC bargaining solution is intuitive. Regardless of which agent makes the TIOLI offer, her objective is to maximize the available surplus of the match. This surplus is generated by transferring more money to the C-type, and it is maximized when the C-type's post-OTC money holdings are $m + \zeta_j = m^*$. However, in order to "afford" this transfer of liquidity, the C-type needs to have enough assets, and the critical level of asset holdings that allows her to acquire $\zeta_j = m^* - m$ depends on who makes the offer. In particular, if the i-type makes the offer that critical level is given by $\bar{d}^i$, $i = \{C, N\}$, where, clearly, $\bar{d}^N > \bar{d}^C$ since if the N-type makes the offer he will ask for more assets to be compensated for $m^* - m$ units of money.

Summing up, if the C-type carries a sufficient amount of assets (defined as $\bar{d}^i$ when the i-type makes the offer), then the money transfer will be optimal, i.e., $\zeta_j = m^* - m$, regardless of who makes the offer, and the asset transfer will satisfy $\chi_j = \bar{d}^i$, where i is the type of agent who makes the offer. On the other hand, if the C-type is constrained by her asset holdings (i.e., if $d_j < \bar{d}^i$ when the i-type makes the offer), then the C-type will give up all her assets, $\chi_j = d_j$, and she will receive a money transfer which is smaller than $m^* - m$ and depends on who makes the offer. More precisely, it satisfies $\zeta_j = d_j$, if the C-type makes the offer, and $\zeta_j = \tilde{\zeta}_j$, if the N-type makes the offer. It is easy to verify that $\tilde{\zeta}_j < d_j$, for all $d_j < \bar{d}^N$, since if the N-type makes the offer she will transfer a lower amount of money to the C-type (for any given amount of assets $d_j < \bar{d}^N$ that she receives).

References

Afonso, G. and R. Lagos (2015). Trade dynamics in the market for federal funds. *Econometrica 83*(1), 263–313.

Alexander, G. J., A. K. Edwards, and M. G. Ferri (2000). The determinants of trading volume of high-yield corporate bonds. *Journal of Financial Markets 3*(2), 177–204.

Amihud, Y. and H. Mendelson (1986). Asset pricing and the bid-ask spread. *Journal of financial Economics 17*(2), 223–249.

Andolfatto, D., A. Berentsen, and C. Waller (2013). Optimal disclosure policy and undue diligence. *Journal of Economic Theory*.

Andolfatto, D. and F. M. Martin (2013). Information disclosure and exchange media. *Review of Economic Dynamics 16*(3), 527–539.

Andolfatto, D., F. M. Martin, and S. Zhang (2015). Rehypothecation and liquidity. *FRB St Louis Paper No. FEDLWP2015-003*.

Andreasen, M. M., J. H. Christensen, K. Cook, and S. Riddell (2016). The tips liquidity premium. *Manuscript, Federal Reserve Bank of San Francisco*.

Arseneau, D. M., D. Rappoport, and A. Vardoulakis (2015). Secondary market liquidity and the optimal capital structure. *Available at SSRN 2594558*.

Berentsen, A., G. Camera, and C. Waller (2007). Money, credit and banking. *Journal of Economic Theory 135*(1), 171–195.

Berentsen, A., S. Huber, and A. Marchesiani (2014). Degreasing the wheels of finance. *International economic review 55*(3), 735–763.

Berentsen, A., S. Huber, and A. Marchesiani (2016). The societal benefit of a financial transaction tax. *European Economic Review 89*, 303–323.

Berentsen, A. and C. Waller (2011). Outside versus inside bonds: A modigliani–miller type result for liquidity constrained economies. *Journal of Economic Theory 146*(5), 1852–1887.

Bethune, Z., B. Sultanum, and N. Trachter (2017). Asset issuance in over-the-counter markets. *Federal Reserve Bank of Richmond Working Paper Series* (17-13).

BlackRock (2014). Corporate bond market structure: The time for reform is now. Technical report.

Branch, W. A., N. Petrosky-Nadeau, and G. Rocheteau (2016). Financial frictions, the housing market, and unemployment. *Journal of Economic Theory 164*, 101–135.

Caramp, N. E. (2017). Sowing the seeds of financial crises: Endogenous asset creation and adverse selection.

Chang, B. and S. Zhang (2015). Endogenous market making and network formation. *Available at SSRN 2600242*.

Das, U. S., M. Polan, and M. G. Papaioannou (2008). *Strategic Considerations for First-Time Sovereign Bond Issuers*. Number 2008-2261. International Monetary Fund Working Paper.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005, November). Over-the-counter markets.

Econometrica 73(6), 1815–1847.

Fernández-Villaverde, J. and D. Sanches (2016). Can currency competition work? Technical report, National Bureau of Economic Research.

Fleming, M. J. (2002). Are larger treasury issues more liquid? evidence from bill reopenings. *Journal of Money, Credit and Banking*, 707–735.

Geromichalos, A. and L. Herrenbrueck (2016). Monetary policy, asset prices, and liquidity in over-the-counter markets. *Journal of Money, Credit, and Banking 48*(1), 35–79.

Geromichalos, A. and L. Herrenbrueck (2017). The liquidity-augmented model of macroeconomic aggregates. Working paper, Simon Fraser University.

Geromichalos, A., L. Herrenbrueck, and S. Lee (2018). Asset safety versus asset liquidity. Working paper, UC Davis.

Geromichalos, A., L. Herrenbrueck, and K. Salyer (2016). A search-theoretic model of the term premium. *Theoretical Economics 11*(3), 897–935.

Geromichalos, A., J. M. Licari, and J. Suárez-Lledó (2007, October). Monetary policy and asset prices. *Review of Economic Dynamics 10*(4), 761–779.

Greenwood, R., S. Hanson, and J. C. Stein (2010). A gap-filling theory of corporate debt maturity choice. *The Journal of Finance 65*(3), 993–1028.

Helwege, J. and L. Wang (2016). Liquidity and price pressure in the corporate bond market: Evidence from mega-bonds.

Herrenbrueck, L. (2019a). Frictional asset markets and the liquidity channel of monetary policy. *Journal of Economic Theory 181*, 82–120.

Herrenbrueck, L. (2019b). Interest rates, moneyness, and the fisher equation. Working paper, Simon Fraser University.

Herrenbrueck, L. and A. Geromichalos (2017). A tractable model of indirect asset liquidity. *Journal of Economic Theory 168*, 252 – 260.

Hotchkiss, E. S. and G. Jostova (2007). Determinants of corporate bond trading: A comprehensive analysis.

Hu, T.-W. and G. Rocheteau (2015). Monetary policy and asset prices: A mechanism design approach. *Journal of Money, Credit and Banking 47*(S2), 39–76.

Huber, S. and J. Kim (2017). On the optimal quantity of liquid bonds. *Journal of Economic Dynamics and Control 79*, 184–200.

Krishnamurthy, A. and A. Vissing-Jorgensen (2012). The aggregate demand for treasury debt. *Journal of Political Economy 120*(2), 233–267.

Lagos, R. (2010, November). Asset prices and liquidity in an exchange economy. *Journal of Monetary Economics 57*(8), 913–930.

Lagos, R. and G. Rocheteau (2008, September). Money and capital as competing media of exchange. *Journal of Economic Theory 142*(1), 247–258.

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

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

Lagos, R., G. Rocheteau, and R. Wright (2017). Liquidity: A new monetarist perspective. *Journal of Economic Literature 55*(2), 371–440.

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

Lagos, R. and S. Zhang (2015). Monetary exchange in over-the-counter markets: A theory of speculative bubbles, the fed model, and self-fulfilling liquidity crises. Technical report, National Bureau of Economic Research.

Lester, B., A. Postlewaite, and R. Wright (2012). Information, liquidity, asset prices, and monetary policy. *The Review of Economic Studies 79*(3), 1209–1238.

Mattesini, F. and E. Nosal (2016). Liquidity and asset prices in a monetary model with otc asset markets. *Journal of Economic Theory 164*, 187–217.

Nosal, E. and G. Rocheteau (2013). Pairwise trade, asset prices, and monetary policy. *Journal of Economic Dynamics and Control 37*(1), 1–17.

Oehmke, M. and A. Zawadowski (2016). The anatomy of the cds market. *The Review of Financial Studies 30*(1), 80–119.

Rocheteau, G. (2011). Payments and liquidity under adverse selection. *Journal of Monetary Economics 58*(3), 191–205.

Rocheteau, G. and A. Rodriguez-Lopez (2014). Liquidity provision, interest rates, and unemployment. *Journal of Monetary Economics 65*, 80–101.

Rust, J. (1985). Stationary equilibrium in a market for durable assets. *Econometrica: Journal of the Econometric Society*, 783–805.

Rust, J. (1986). When is it optimal to kill off the market for used durable goods? *Econometrica: Journal of the Econometric Society*, 65–86.

Song, Z. and H. Zhu (2018). Quantitative easing auctions of treasury bonds. *Journal of Financial Economics 128*(1), 103–124.

Üslü, S. (2015). Pricing and liquidity in decentralized asset markets. Working paper, UCLA.

Vayanos, D. and J. Wang (2012). Market liquidity–theory and empirical evidence. Technical report, National Bureau of Economic Research.

Vayanos, D. and T. Wang (2007). Search and endogenous concentration of liquidity in asset markets. *Journal of Economic Theory 136*(1), 66–104.

Vayanos, D. and P.-o. Weill (2008). A search-based theory of the on-the-run phenomenon. *The Journal of Finance 63*(3), 1361–1398.

Venkateswaran, V. and R. Wright (2013). Pledgability and liquidity: a new monetarist model of financial and macroeconomic activity. Working paper, NBER.

Weill, P.-O. (2007). Leaning against the wind. *The Review of Economic Studies 74*(4), 1329–1354.

Weill, P.-O. (2008). Liquidity premia in dynamic bargaining markets. *Journal of Economic Theory 140*(1), 66–96.