A Search-Theoretic Model of the Term Premium

Athanasios Geromichalos
University of California - Davis

Lucas Herrenbrueck
Simon Fraser University

Kevin Salyer
University of California - Davis

This Version: September 2015

ABSTRACT

A consistent empirical feature of bond yields is that term premia are, on average, positive. The majority of theoretical explanations for this observation have viewed the term premia through the lens of the consumption based capital asset pricing model. In contrast, we harken to an older empirical literature which attributes the term premium to the idea that short maturity bonds are inherently more liquid. The goal of this paper is to provide a theoretical justification of this concept. To that end, we employ a monetary-search model extended to include assets of different maturities. Short term assets mature in time to take advantage of random consumption opportunities. Long term assets cannot be used directly to purchase consumption, but agents may liquidate them in a secondary asset market characterized by search and bargaining frictions. Our model delivers three results that are consistent with empirical facts. First, long term assets have higher rates of return in steady state to compensate agents for their relative lack of liquidity. Second, since the difference in the yield of short and long term assets reflects asset market frictions, our model predicts a steeper yield curve for assets that trade in less liquid secondary markets. Third, our model predicts that freshly issued ("on-the-run") assets will sell at higher prices than previously issued ("off-the-run") assets that mature in nearby dates, because sellers of the latter have a more urgent need for liquidity.

JEL Classification: E31, E43, E52, G12

Keywords: monetary-search models, liquidity, over-the-counter markets, yield curve

Email: ageromich@ucdavis.edu, herrenbrueck@sfu.ca, kdsalyer@ucdavis.edu.
We are grateful to Borağan Aruoba, Saki Bigio, Jeffrey Campbell, Francesca Carapella, Jonathan Chiu, Darrell Duffie, Ricardo Lagos, Benjamin Lester, Guillaume Rocheteau, Peter Rupert, Ina Simonovska, Alberto Trejos, Venky Venkateswaran, and Randall Wright for useful comments and suggestions.

1 Introduction

An established feature of bond yields is that, on average, investors in long term bonds receive a higher return than investors in similar (e.g. same default risk and tax treatments) short term bonds over a given holding period. As is well known, this observation, which implies the presence of positive term premia, is inconsistent with the expectations hypothesis of the term structure of interest rates. A traditional explanation of such term premia, popular in the empirical finance literature, is based on the idea that short term bonds are inherently more liquid. For instance, Cochrane (1999) compares average returns on bonds of different maturities and states that the "increase in returns for long term bonds, equivalent to [an] upward slope in the yield curve, is usually excused as a small 'liquidity premium'". The same idea is also presented as one of the main explanations for an upward sloping yield curve in most undergraduate textbook discussions of the term structure (e.g. Mishkin (2007)). The goal of this paper is to develop a framework that formalizes the notion of asset liquidity, and to use this framework to provide a rigorous foundation for the idea that term premia can, at least in part, be attributed to market frictions and liquidity differentials.

Our explanation is quite intuitive: in a world with uncertain consumption expenditures (e.g. health costs, home repairs, etc.), there may be times when some of the illiquid assets held by households must be converted to liquid assets in order to finance unexpected consumption expenditures. However, in our model, the asset market in which this conversion takes place is not a perfectly competitive Walrasian market but, instead, has the features of an over-the-counter (OTC) market characterized by search and bargaining, such as that described in Duffie, Gârleanu, and Pedersen (2005). Moreover, to provide a precise definition of asset liquidity, we employ the search-theoretic framework of Lagos and Wright (2005), in which a subset of goods markets are decentralized and anonymous, so that a medium of exchange emerges naturally. In this environment, an asset's value is determined both by fundamentals (as in a standard asset pricing model) and by the ease with which the asset can be used for consumption purchases; i.e. its liquidity properties.

In our model, all assets mature before the decentralized goods market opens (on their respective maturity date). Consequently, short term assets are endogenously closer substitutes to money simply because these assets mature in time to take advantage of random consumption opportunities. This, in turn, implies that, in equilibrium, short term assets are priced above their fundamental value: they carry a *liquidity premium*. Longer term assets, i.e. those that have not matured, cannot serve as a means of payment. However, agents who carry these assets but have a demand for liquidity can visit an OTC financial market where they can exchange them for liquid assets with agents who do not wish to consume in the current period. We show that, although long term assets cannot be used directly to purchase goods, in equilibrium, they can carry an *indirect* liquidity premium, which reflects their ability to help agents avoid the cost of

carrying liquid assets, a cost which is strictly positive in all monetary equilibria.

The main result of the paper is that, if asset supply is not too large, long maturity assets will sell at a discount (i.e. "haircut") relative to short maturity assets; that is, investors must be compensated for holding the relatively illiquid long maturity assets so that a positive term premium emerges in equilibrium. We illustrate that this term premium is closely linked to the search and bargaining frictions characterizing the OTC asset market. In particular, we demonstrate that the only way to obtain a zero term premium is if the agents who have an opportunity to consume are guaranteed to trade in the OTC market, *and* if they can extract the whole surplus generated from OTC trade. It is important to highlight that relative asset scarcity is a necessary condition for the existence of a term premium: if asset supply is so large that maturing assets (together with real money balances) can cover the total liquidity needs of the economy, then assets will be priced at their fundamental value, implying a flat yield curve.[1]

The model also allows us to study the effect of inflation on asset prices. When (anticipated) inflation increases, the price of short term assets, which are effectively substitutes to money, also rises. Since both money and short term assets are now more costly to hold, agents also value longer term assets more, since, as explained earlier, the latter can help agents avoid the inflation tax. Hence, a higher inflation tends to increase (decrease) the equilibrium price (yield) of assets of all maturities. However, since short term assets are closer substitutes to money, the term premium, i.e. the slope of the yield curve, is typically increasing in inflation.

One of the key insights of our model is that the issue price of long maturity assets is crucially (and positively) affected by the liquidity of the secondary asset market, i.e. how easy it is for agents to liquidate these assets in the OTC market. To highlight the importance of this channel, we extend the baseline model to include a second set of assets that only differ from the original ones in that they cannot be traded in secondary markets (i.e. agents have to hold them to maturity). We show that the issue price of long maturities will be higher for the assets that can be traded in secondary financial markets, thus reflecting a liquidity premium. Krishnamurthy and Vissing-Jorgensen (2012) compare the yields on 6-month FDIC-insured certificates of deposit (CDs) and 6-month treasury bills over the 1984-2008 period and provide direct evidence in support of our finding.[2] In particular, they report that the spread was 2.3 percentage points on average which they attribute to the higher liquidity of T-bills. Moreover, the authors report that the spread between the yields of the two assets is negatively related to the supply of T-bills, a result that we also obtain in our theoretical analysis. The prediction that bond yields are influenced by the liquidity of secondary markets is also consistent with Gürkaynak, Sack,

[1] More formally, let ψ_i and r_i be the price and the interest rate of a pure discount bond that yields one unit of consumption in i periods. As of period t, the "fundamental value" of the bond with maturity i is β^i, where $\beta \in (0,1)$ is the discount rate. Furthermore, from the undergraduate textbook, the formula that links the price and interest rate of such bonds is $\psi_i = (1+r_i)^{-i}$, implying that if assets are priced at "fundamental", then $r_i = 1/\beta - 1$, for all i, or, equivalently, the yield curve is flat.

[2] Notice that the assets under consideration have the same maturity and the same default risk (they are default-free). However, unlike T-bills, CDs have to be held to maturity.

and Wright's (2010) analysis of the yield curve for inflation-indexed Treasury debt (i.e. TIPS). In particular, they demonstrate that TIPS yields have, in general, fallen as market liquidity (measured by trading volume) in the TIPS market has increased.

Our framework also allows us to compare the price of freshly issued (on-the-run) short term assets with the price of older assets (off-the-run) which mature on the same date. Conventional wisdom suggests that the yields on assets with identical streams of dividends should be equal. However, Warga (1992) documents that the return of an off-the-run portfolio exceeds, on average, the return of an on-the-run portfolio with similar duration. Our model is consistent with this observation. Intuitively, in our analysis, the sellers of off-the-run assets are agents who received an opportunity to consume, and who are desperate for liquidity and, thus, more willing to sell assets at a lower price. Vayanos and Weill (2008) also provide a theoretical explanation of the "on-the-run phenomenon", by building a model where on the on-the-run bonds are more liquid (i.e. easier to sell) because they constitute better collateral for borrowing in the repo market. Although the two models are quite different, they share a common feature which is essential for their ability to capture the on-the-run phenomenon: the assumption that asset trade takes place in OTC markets.

1.1 Related Literature

The literature on the term structure is vast but the profession is fortunate to have several excellent survey articles. In particular, Gürkaynak and Wright (2012) provide a nice overview of the testable implications of the expectations hypothesis and the lack of empirical support for them. Their review is from a macroeconomic perspective, as is our analysis. For a discussion of analyses of the term structure from a finance tradition, the reader is referred to Singleton (2009) and Piazzesi (2010).

Given the failure of the expectations hypothesis, the challenge to economists is to define and quantify the nature of risks and market frictions associated with the purchase of long term bonds. The majority of responses to this challenge have identified the term premium as a risk premium and employed the consumption based CAPM in their analyses. In such models, the qualitative nature of asset risk is characterized by the covariance between investors' stochastic discount factors and asset returns. Early papers (Backus, Gregory, and Zin (1989) and Salyer (1990)) demonstrated, however, that due to the autocorrelation properties of inflation, a standard intertemporal asset pricing model would produce counterfactual negative term premia for nominally denominated bonds. More recently, Piazzesi and Schneider (2007) combined Epstein-Zin preferences with a richer stochastic model of inflation and consumption growth and demonstrated that these features can indeed produce positive, time-varying nominal term premia consistent with observation.

While certainly insightful, we do not view the above explanation as wholly satisfactory. For one, positive term premia over holding horizons as short as one quarter are routinely observed

(Backus, Gregory, and Zin (1989)). It is difficult to see this as a response to changes in long run risk as explained by Piazzesi and Schneider (2007). As an alternative explanation, one that we view as complementary rather than competing, we present a model of the term premium which is based upon the inherent liquidity differentials between bonds of different maturities.

Our paper is related to the recent work by Williamson (2013), who also studies the term premium within a monetary-search model in which assets help agents carry out transactions in markets with imperfect credit (by serving as collateral). The author assumes that short term assets are more "pledgeable", i.e. they are better facilitators of trade in the decentralized goods market. This assumption is crucial for the existence of an upward sloping yield curve, which, in turn, is crucial for the unconventional monetary policy considered in the paper (i.e. quantitative easing) to be effective. Our paper shows that an upward sloping yield curve arises even if short term assets are not given an exogenous liquidity advantage over long term assets, other than the fact that they mature earlier.

Our work is also related to a number of papers where the existence of uninsured idiosyncratic risks and borrowing constraints can generate a preference for short over long assets; e.g. Heaton and Lucas (1992) and Challe, Le Grand, and Ragot (2013). In these papers, a bad income shock may force agents to sell assets before maturity, and the payoff of selling is uncertain due to aggregate risk.[3] While the fact that agents may find themselves in need of selling long assets before maturity is a feature that we share with these papers, there are also important differences. First, in our model there is no aggregate uncertainty, which may make our story more relevant to (the yield curve for) assets such as government bonds and to short horizons. Moreover, in our model, purchasing long term bonds is associated with certain costs that are directly linked to the degree of (il)liquidity of secondary asset markets (sometimes referred to as market "microstructure"). As a result, our paper has a number of empirically supported results other than the upward sloping yield curve. For instance, our model predicts that the yield on long term assets will be higher for assets that trade in less liquid secondary markets, and it provides an intuitive explanation for the on versus off-the run phenomenon.

This paper is also related to a growing literature that studies the liquidity properties of assets other than money. Examples include Geromichalos, Licari, and Suarez-Lledo (2007), Lagos and Rocheteau (2008), Lagos (2011), Lester, Postlewaite, and Wright (2012), Nosal and Rocheteau (2013), Jacquet and Tan (2012), Andolfatto and Martin (2013), Williamson (2012), and Andolfatto, Berentsen, and Waller (2014). Some recent papers exploit the idea that asset prices can carry liquidity premia to offer a new perspective for looking at long-standing asset pricing-related puzzles. Examples include Lagos (2010) (equity premium and risk-free rate puzzles) and Geromichalos and Simonovska (2014) (asset home bias puzzle). Our paper is conceptually related to this literature, since it demonstrates how asset liquidity can help rationalize the

[3] In that sense, these papers are conceptually related to the aforementioned literature (e.g. Backus *et al* (1989), in which agents command higher returns on long term assets because they dislike future interest rate risk.

term premium puzzle. Finally, a number of papers, such as Boel and Camera (2006), Berentsen, Camera, and Waller (2007), and Berentsen, Huber, and Marchesiani (2014) explore the idea that assets may carry liquidity premia because they allow agents to rebalance their money holdings after a consumption opportunity arises. The present paper is uniquely identified from these papers, in that we model explicitly the frictions present in the secondary asset market, and in that we consider assets of different maturities in order to focus on the term premium.

The rest of the paper is organized as follows. In Section 2, we describe the physical environment. In Section 3, we characterize equilibrium in a simple version of the model with two maturities and without money. Section 4 shows how the main results of Section 3 can be generalized in an environment with money, and with assets of any number of maturities. Section 5 offers some concluding comments.

2 The Model

Before presenting a detailed description of the economy, we first describe the basic setup in order to highlight the critical features of the model. Our framework generalizes that of Geromichalos and Herrenbrueck (2012) to include assets with different maturities. Specifically, we employ an infinite horizon, discrete-time economy in which each period is divided into three subperiods. The subperiods are identified by their markets. At the beginning of the period, a financial market opens in which assets with different maturities can be traded. This market resembles the over-the-counter market of Duffie, Gârleanu, and Pedersen (2005), and we refer to it as the OTC market. In the second subperiod, agents meet in decentralized markets characterized by anonymous, bilateral trades, as in Lagos and Wright (2005). We refer to it as the LW market. In the final subperiod, trade of newly available assets and goods takes place in a centralized (i.e. Walrasian) market, which we refer to as the CM. More details of these markets are given below.

The economy contains two types of infinitely-lived agents, buyers and sellers, defined by their actions in the LW market. The measure of buyers is normalized to one. At the beginning of each period, a fraction $\ell < 1$ of buyers learn that they have an opportunity to purchase the good sold in the LW market; we refer to them as the C-type buyers. The remaining measure of buyers, $1 - \ell$, denoted N-types, do not purchase goods in the LW market of the current period (i.e. they will be inactive buyers in the current period). To keep the analysis simple, we assume that all C-types match with a seller in the LW market, and we accordingly set the measure of sellers to equal ℓ.

The role of the markets is as follows. Buyers, who are the agents that make all the interesting decisions in our model, leave the CM in the previous period and then find out whether they will be consuming in the LW market. Since trade is anonymous in that market, C-type buyers need a medium of exchange to finance their purchases. The OTC market is strategically placed

before the LW market opens, but after the uncertainty regarding consumption in the LW market has been resolved, in order to allow agents who might be short of "liquidity" (defined below) to exchange illiquid assets for liquid assets.

Returning to the structure of the model, all agents discount the future between periods (but not subperiods) at rate $\beta \in (0,1)$. Buyers consume in the second and third subperiods and supply labor in the third subperiod. Their preferences for consumption and labor within a period are given by $\mathcal{U}(X,H,q)$, where X,H represent consumption and labor in the CM, respectively, and q consumption in the LW market. Sellers consume only in the CM and produce in both the CM and the LW market. Their preferences are given by $\mathcal{V}(X,H,h)$, where X,H are as above, and h stands for hours worked in the LW market. The sellers' production technology of LW good is given by $q = h$. Following Lagos and Wright (2005), we adopt the functional forms $\mathcal{U}(X,H,q) = U(X) - H + u(q)$ and $\mathcal{V}(X,H,h) = U(X) - H - c(h)$. Assume that u, U are twice continuously differentiable with $u(0) = 0$, $u' > 0$, $u'(0) = \infty$, $u'(\infty) = 0$, $U' > 0$, $u'' < 0$, and $U'' \leq 0$. For simplicity, we set $c(h) = h$, but this is not crucial for any results. Let q^* denote the optimal level of output in any LW market meeting, i.e. $q^* \equiv \{q : u'(q^*) = 1\}$. Also, there exists $X^* \in (0,\infty)$ such that $U'(X^*) = 1$, with $U(X^*) > X^*$.

In the third subperiod, all agents consume and produce a general good which we refer to as fruit. The supply of this good comes from two sources: labor supplied by agents, and the output (i.e. dividend) of assets maturing that period. Agents have access to a technology that transforms one unit of labor into one unit of the fruit. Each period, the economy is endowed with a set of trees, as in Lucas (1978), that deliver a real dividend (i.e. fruit) at different dates (maturities). Each share of a tree of maturity $i \in \{1,...,N\}$ purchased in period t, delivers 1 unit of fruit in period $t+i$. For reasons that will become clear later, we assume that the fruit is delivered before the LW market opens. Agents can store the fruit at no cost between the second and the third subperiod (when they consume it), but the fruit is perishable between time periods. Agents can purchase any amount of shares of a tree of maturity i at the ongoing real market price $\psi_{i,t}$. The supply of trees that mature in i periods is denoted by $A_i > 0$, and it is fixed over time. Importantly, by this definition the supply includes newly issued i-period trees and older trees that mature at the same date.

In addition to the trees, there also exists an asset called "money" in this economy. Money has no intrinsic value, but it is infinitely lived, storable, divisible, and recognizable by all agents. Hence, it can serve as a medium of exchange in the LW market and help bypass the frictions created by anonymity and the lack of a double coincidence of wants. The market price of money in terms of fruit is denoted by φ_t. Its supply is controlled by a 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.

The anonymous, bilateral exchanges in the LW market are characterized by take-it-or-leave-it offers made by the buyer to the seller. Due to the anonymity in that market, exchange has

to be *quid pro quo*, and the only objects that can serve as means of payment are money and the yield of assets that have already matured; critically, claims to trees that mature in future periods can not be used for payment. These assumptions are discussed in detail in Section 2.1 below.

If a C-type agent finds herself in need of additional liquidity, she can visit the OTC market and search for a trading partner (an N-type) who might hold some liquid assets (i.e. money and trees that pay out in the current period) that she will not use in this period's LW market. Hence, gains from trade can be generated by C-types selling a portfolio of long term assets (i.e. assets that do not mature in the current period) in exchange for assets that mature in the current period. We assume that a matching function, $f(\ell, 1-\ell) \leq \min\{\ell, 1-\ell\}$, brings together C-types and N-types. The function f is homogeneous of degree one and increasing in both arguments. Within each match, the terms of trade are determined through proportional bargaining, following Kalai (1977), and the C-type's bargaining power is given by $\lambda \in (0, 1)$.

Figure 1 summarizes the timing of events in the model, for the case with $N = 2$. For instance, consider a buyer who has just left the CM of period $t-1$ (denoted by CM_{t-1}) and has found out that in period t she is a C-type. This agent possesses liquidity that comes from three potential sources: a) assets of maturity 1 purchased in $t-1$ (indicated by the arrow between CM_{t-1} and the beginning of LW_t); b) assets of maturity 2 purchased in $t-2$ (indicated by the dashed arrow that points to beginning of LW_t); and c) money purchased in $t-1$ (not indicated in the figure, for simplicity). The agent might also hold some assets of maturity 2, purchased in $t-1$ (indicated by the arrow between CM_{t-1} and the beginning of LW_{t+1}). These assets will only mature in $t+1$, but if the agent finds herself in need for additional liquidity in period t, she can visit the OTC market, search for a trading partner (i.e. an N-type), and exchange these assets for a portfolio of money and assets that are about to mature.

It should be pointed out that our analysis treats the assets of various maturities symmetrically: all assets deliver one unit of the numéraire at the beginning of the LW market, but some do so in earlier periods than others. Hence, any differential in the yield of alternative assets is solely due to differences in the date of their maturity.

Throughout the paper we focus on steady-state equilibria, and most of the equilibrium analysis is carried out with respect to the asset prices ψ_i, $i = 1, ..., N$. When we wish to make statements regarding the interest rate of the various assets, we use the standard (textbook) formula that links the price and interest rate of an asset. In particular, we have

$$\psi_i = (1 + r_i)^{-i}, \quad \text{for all } i = 1, ..., N. \tag{1}$$

2.1 Discussion of the Physical Environment

The assumption that the yield of assets can serve as a means of payment aims to capture the simple idea that assets that have matured are *as good (liquid) as money*, while allowing us to

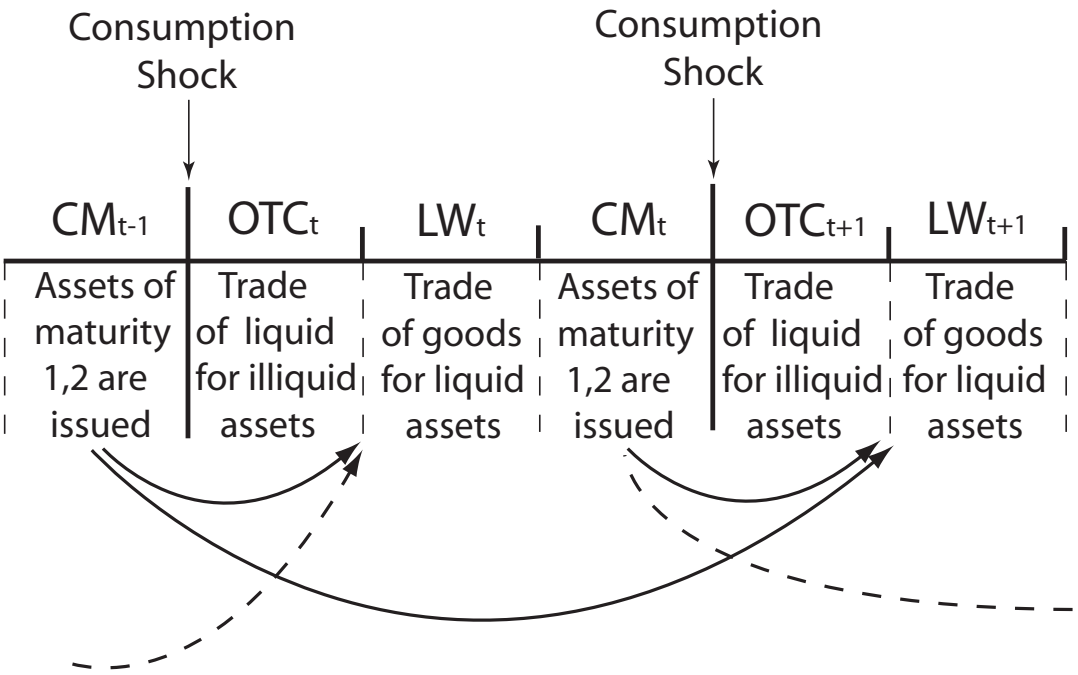

Figure 1: Timing of events in a model with $N = 2$.

work with real rather than nominal assets, as is standard in the monetary-search and theoretical finance literatures. Employing real assets also clarifies that an upward sloping yield curve is not (only) a nominal or inflation-risk phenomenon.[4] An alternative interpretation of this assumption is that maturing assets deliver inflation-indexed money rather than actual fruit. In the accompanying web appendix, we present a version of the model with nominal assets and demonstrate that the main results of the paper remain unaltered. Another crucial assumption is that assets deliver their yield (dividend in the main body of the paper or cash in the web appendix) before the LW market opens. This assumption aims to capture the simple and intuitive idea that short term assets become liquid just when the agent needs them.

In this paper, we take as given that assets (that mature in future periods) themselves cannot be used as media of exchange and study asset prices subject to this restriction on the physical environment. However, there might be deeper reasons why agents choose to not use financial assets in order to purchase consumption goods. For example, Rocheteau (2011) and Lester *et*

[4] This paper aims to study how the various assets' prices are shaped by their respective maturities and the financial market frictions, and adopting real assets with an exogenous supply is the most neat way to carry out this exercise. But let us note two things. First, while our paper does not provide a theory of endogenous determination of A_i, $i = 1, ..., N$, (because this is not the question we are after), it also does not place any strong restrictions on the value of asset supplies necessary for the main results. In fact, the main results (i.e. Results 1-3 in Section 3) are valid as long as the supply of maturing assets, A_1, is not sufficient to cover all the liquidity needs of the economy (see Section 3.6 for details). Second, there is evidence indicating that a fixed asset supply assumption might actually be empirically plausible. For instance, Garbade (2007) reports that the Treasury abandoned "tactical" debt management and, instead, adopted a policy of issuing notes and bonds on a "regular and predictable" schedule, because the latter approach was credited with "reducing market uncertainty, facilitating investor planning, and lowering the Treasury's borrowing costs".

al (2012) both consider environments that do not place any restrictions on which objects can serve as media of exchange. They show that, if there is asymmetric information regarding the future returns of financial assets, fiat money (or, in our case, the physical yield of assets that have already matured) will arise endogenously as a superior medium of exchange in bilateral meetings. A general reason that might give money or physical goods an advantage over assets as a medium of exchange is recognizability. A seller might be reluctant to accept an asset (i.e. a claim to future dividends) as a medium of exchange, because this asset might not even be a tangible object (e.g. it could be an entry in a computer).[5]

In our model, all real assets are first traded (issued) in Walrasian markets. This is a methodological innovation, due to Lagos and Wright (2005), which (together with quasi linearity of preferences) gives rise to degenerate asset holding distributions and ensures tractability. This setup is convenient and, to some extend, realistic. Many assets that are eventually traded in OTC secondary markets are indeed issued in primary markets with more competitive characteristics. For instance, US Treasury Bills are issued through single-price auctions, in order "to minimize the government's costs [...] by promoting broad, competitive bidding" (Garbade and Ingber (2005)). It should be pointed out that *only* newly issued assets are traded in the CM. Agents who wish to sell assets that mature in future periods can only do so in the OTC market. In other words, all secondary asset trade is OTC.[6]

The assumption that asset liquidation takes place in a frictional OTC market is not only crucial for our theory but also empirically relevant. For instance, Green (1993) reports that the secondary market for municipal bonds is characterized by significant trading delays. OTC markets for treasuries are considered more liquid because they are characterized by small trading delays and bid-ask spreads, however, one needs to note that trading delays are small because trade is intermediated by market-makers or dealers. In fact, the very existence of dealers in these markets serves as evidence that they are not competitive. Dealers arise naturally in these markets because of their ability to bypass the frictions that make *direct* trade among investors difficult. Bid-ask spread for US bonds ranges from roughly 0.8 basis points for 2-year notes to 1.5 basis points for 10-year bonds (Engle, Fleming, Ghysels, and Nguyen (2012)); these are small compared to other markets but, when multiplied by the enormous volume of trade (in the range of hundreds of billions per day), they represent significant absolute trading costs. Even for the next most liquid asset class, high-quality corporate bonds, bid-ask spreads are much

[5] Another justification for the coexistence of money and higher-return assets is provided by Hu and Rocheteau (2013), who show that this coexistence is a feature of an optimal mechanism.

[6] We think of the CM as the market where assets are first issued/auctioned. Hence, allowing agents to sell assets off-the-run in the CM would be like allowing them to show up at the auction of new assets and try to sell their own assets alongside with the issuer (e.g. the Treasury or a municipal government), which we think would be unreasonable. That said, it should also be clarified that the assumption in question does not affect the analysis in any major way. The only consequence is that we have to rule out a (small) subset of parameter values where it is impossible to characterize equilibrium. For more details, see the discussion in Section 3.4.

larger: on the order of 8 basis points in 2007 and almost twice as high in recent years.[7]

3 Equilibrium in the Model with $N = 2$ and no Money

In this section, we focus on the case of two maturities in an economy without money. This version of the model conveys all the economic insights that we wish to highlight, and delivers the most important results of the paper. The full model with money is solved in the web appendix to this paper, but we will summarize the additional results concerning the effect of money on asset prices and the term premium in Section 4.1.

3.1 Value Functions

We begin with the description of the value functions in the CM. For a typical buyer, the state variables are the following. First, the dividend, z, that she received earlier in the period, i.e. before the LW market opened, and she did not spend in that market. The amount of real balances z could have been delivered either from long term assets issued two periods ago, or from short term assets issued in the last period. Second, the units of assets of maturity $N = 2$, a_2, that she bought in the previous period, and which will mature in the forthcoming period. The Bellman equation is given by

$$W(z, a_2) = \max_{X,H,\hat{a}_1,\hat{a}_2} \{U(X) - H + \beta \mathbb{E}\{\Omega^i(\hat{a}_1, \hat{a}_2)\}\}$$
$$\text{s.t. } X + \psi_1(\hat{a}_1 - a_2) + \psi_2 \hat{a}_2 = H + z,$$

and subject to $\hat{a}_1 - a_2 \geq 0$. In the last expression, variables with hats denote next period's choices, and the term $\mathbb{E}$ denotes the expectations operator. The function Ω^i represents the value function in the OTC market for a buyer of type $i = \{C, N\}$, described in more detail below. It is important to highlight that we have defined $\hat{a}_1$ as the amount of *all* assets that mature in the next period (which is analogous to our definition of the supply of assets that mature in the next period). Hence, the amount of newly issued short term assets purchased by the agent is $\hat{a}_1 - a_2$, and we require $\hat{a}_1 - a_2 \geq 0$. This constraint simply enforces the assumption that agents cannot sell off-the-run short term asset in the CM (see the discussion in footnotes 6 and section 3.4), and for the rest of the paper, we will focus only on equilibria where this constraint does not bind.

Some observations are in order. First, it can be easily verified that, at the optimum, $X = X^*$.

[7] This evidence is based on a Barclays 2015 report on financial liquidity, titled "The decline in financial market liquidity". The report is available upon request.

Using this fact and replacing H from the budget constraint into W yields

$$\begin{aligned} W(z,a_2) &= U(X^*) - X^* + z + \psi_1 a_2 \\ &+ \max_{\hat{a}_1,\hat{a}_2} \left\{ -\psi_1 \hat{a}_1 - \psi_2 \hat{a}_2 + \beta \mathbb{E}\left\{ \Omega^i(\hat{a}_1,\hat{a}_2) \right\} \right\}. \end{aligned} \tag{2}$$

A standard feature of models that build on Lagos and Wright (2005) is that the optimal choice of the agent does not depend on the current state (due to the quasi-linearity of $\mathcal{U}$). This is also true here, with the exception that the range of admissible choices for $\hat{a}_1$ is restricted by the state variable a_2. Moreover, as is standard in this types of models, the CM value function is linear in z (again, as long as the constraint $\hat{a}_1 - a_2 \geq 0$ does not bind for any agent, which will be the case in all the equilibria we consider in this paper). We collect all the terms in (2) that do not depend on the state variables, and we write

$$W(z,a_2) = \Lambda + z + \psi_1 a_2, \tag{3}$$

where the definition of Λ is obvious.

Next, consider a seller's value function in the CM. It is well-known that in monetary models where the identity of agents (as buyers or sellers) is fixed over time, sellers will typically not leave the CM with a positive amount of asset holdings.[8] When a seller enters the CM, she will typically hold some real balances, z, that she received as payment during trade in the preceding LW market. Also, the seller does not visit the OTC market. Therefore, it can be easily shown that her CM value function is given by

$$W^S(z) = \Lambda^S + z, \tag{4}$$

where $\Lambda^S \equiv U(X^*) - X^* + \beta V^S$, and V^S denotes the seller's value function in next period's LW market, to be discussed below.

Consider now the value functions in the LW market. Let q denote the quantity of special good produced, and π the real balances that change hands during trade in the LW market. These terms will be determined in Section 3.2. The LW value function for a buyer who enters that market with portfolio (z,a_2) is given by

$$V(z,a_2) = u(q) + W(z-\pi,a_2), \tag{5}$$

[8] The intuition behind this result is simple. In monetary models, assets will, in general, be priced above the "fundamental value", reflecting liquidity premia. Agents who know with certainty that they will not have an opportunity to consume in the forthcoming LW market (just like our sellers here) will not be willing to pay such premia. Here we take this result as given (for a detailed discussion, see Rocheteau and Wright (2005)).

and the LW value function for a seller (who enters with no assets) is given by

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

Finally, consider the value functions in the OTC market. After leaving the CM, and before the OTC market opens, buyers learn whether they will have a chance to access this period's LW market (C-types) or not (N-types). This chance will occur with probability $\ell \in (0,1)$. The expected value for the typical buyer, before she enters the OTC market, is given by

$$\mathbb{E}\left\{\Omega^i(m,a_1,a_2)\right\} = \ell\,\Omega^C(m,a_1,a_2) + (1-\ell)\,\Omega^N(m,a_1,a_2). \tag{6}$$

In the OTC market, C-type buyers, who may want additional liquid assets, are matched with N-type buyers, who may hold liquid assets that they will not use in the current period. Hence, trade in the OTC involves C-types giving up long term assets for short term assets. Given the matching function $f(\ell, 1-\ell)$, define the matching probabilities for C-types and N-types as $\alpha_{_C} \equiv f(\ell,1-\ell)/\ell$ and $\alpha_{_N} \equiv f(\ell,1-\ell)/(1-\ell)$, respectively. Let χ denote the units of long term assets that the C-type transfers to the N-type, and ζ the units of liquid assets that the C-type receives in return. These terms will be determined in Section 3.2. Then,

$$\Omega^C(m,a_1,a_2) = \alpha_{_C}\, V\left(z+\zeta, a_2-\chi\right) + (1-\alpha_{_C})V(z,a_2), \tag{7}$$

$$\Omega^N(m,a_1,a_2) = \alpha_{_N}\, W\left(z-\zeta, a_2+\chi\right) + (1-\alpha_{_N})\, W(z,a_2). \tag{8}$$

Notice that N-type buyers proceed directly to the CM.

We now proceed to the description of the terms of trade in the LW and the OTC markets.

3.2 Bargaining in the LW and OTC Markets

Consider a meeting between a C-type buyer with real balances z and long term assets a_2, and a seller who, as we have argued, holds no real balances or assets as she enters the LW subperiod. The two parties bargain over a quantity q, to be produced by the seller, and a real payment π, to be made to the seller. The buyer makes a take-it-or-leave-it offer, maximizing her surplus subject to the seller's participation constraint. The bargaining problem can be described by

$$\max_{\pi,q} \left\{u(q) + W(z-\pi,a_2) - W(z,a_2)\right\},$$

subject to $-q + W^S(\pi) - W^S(0) = 0$ and the constraint $\pi \leq z$. Taking advantage of the linearity of W, W^S (equations (3) and (4)), allows us to simplify the bargaining problem to

$$\max_{\pi,q} \{u(q) - \pi\},$$

subject to $q = \pi$ and $\pi \leq z$. The solution to this problem is as follows.

Lemma 1. *The solution to the bargaining problem is given by* $q(z) = \pi(z) = \min\{q^*, z\}$.

Proof. This result is very standard in the literature, hence, the proof is omitted. For a detailed proof see Geromichalos *et al* (2007) or Lester *et al* (2012). □

The solution to the bargaining problem is very intuitive. The only variable that affects the solution is the buyer's real balances. As long as the buyer carries q^* or more, the first-best quantity q^* will always be exchanged. On the other hand, if $z < q^*$, the buyer does not have enough liquidity to induce the seller to produce q^*. In this case, the buyer will give up all her real balances, $\pi(z) = z$, and the seller will produce the quantity of good that satisfies her participation constraint, that is $q = \pi(z) = z$.

Turning to the OTC market, consider a meeting between a C-type with portfolio (z, a_2), and an N-type with portfolio $(\tilde{z}, \tilde{a_2})$, where we again identify a_1 with z to recognize that one unit of short term assets delivers one unit of the dividend. Let χ denote the long term assets that the C-type transfers to the N-type, and let ζ represent the liquid assets received by the C-type. Also, let S^i, $i = \{C, N\}$, denote the surplus of type i, and $\lambda \in [0, 1]$ the bargaining power of the C-type. With proportional bargaining, the objective is to choose χ, ζ in order to maximize S^C, subject to: a) the constraint that the ratio S^C/S^N should be equal to the ratio $\lambda/(1-\lambda)$, and b) the feasibility constraints $\chi \leq a_2$ and $\zeta \leq \tilde{z}$. The terms S^i are given by:[9]

$$S^C \equiv V(z + \zeta, a_2 - \chi) - V(z, a_2),$$
$$S^N \equiv W(\tilde{z} - \zeta, \tilde{a}_2 + \chi) - W(\tilde{z}, \tilde{a}_2).$$

Substituting for W, V from (3) and (5) in the expression above, and exploiting Lemma 1 (the LW bargaining solution) allows us to write[10]

$$S^C = u(z + \zeta) - u(z) - \psi_1 \chi,$$
$$S^N = \psi_1 \chi - \zeta.$$

[9] Since the C-type has a consumption opportunity, she will proceed to the LW market with an additional ζ units of real balances, but also with her long term asset holdings reduced by the amount χ. The N-type will proceed directly to the CM with less money and short term assets, but with more long term assets.

[10] We focus on the case where $z < q^*$, since, if the opposite is true, the C-type is carrying the maximum possible liquidity, and no trade in the OTC can generate a positive surplus. Moreover, we restrict attention to OTC trades that involve a transfer of real balances $\zeta \leq q^* - z$, i.e. we assume that the C-type will never acquire more real balances than she needs in order to attain the first-best quantity q^*.

Therefore, the OTC bargaining problem can be written as

$$\max_{\chi,\zeta} \Big\{ u\left(z+\zeta\right) - u\left(z\right) - \psi_1 \chi \Big\}, \tag{9}$$

$$\text{s.t. } u\left(z+\zeta\right) - u\left(z\right) - \psi_1 \chi = \frac{\lambda}{1-\lambda}\left(\psi_1 \chi - \zeta\right), \tag{10}$$

and $\chi \leq a_2$, $\zeta \leq \tilde{z}$.[11] The solution to this problem is described in the following lemma.

Lemma 2. *Consider a meeting in the OTC market between a C-type and an N-type with portfolios (z, a_2) and $(\tilde{z}, \tilde{a}_2)$, respectively, and define the cutoff level of long term asset holdings*

$$\bar{a}(z,\tilde{z}) \equiv \frac{1}{\psi_1}\Big\{ (1-\lambda)\left[u\left(\min\{z+\tilde{z}, q^*\}\right) - u(z)\right] + \lambda \min\{q^* - z, \tilde{z}\} \Big\}. \tag{11}$$

Then, the solution to the bargaining problem is given by

$$\chi(z,\tilde{z},a_2) = \begin{cases} \bar{a}(z,\tilde{z}), & \text{if } a_2 \geq \bar{a}(z,\tilde{z}), \\ a_2, & \text{if } a_2 < \bar{a}(z,\tilde{z}). \end{cases} \tag{12}$$

$$\zeta(z,\tilde{z},a_2) = \begin{cases} \min\{q^* - z, \tilde{z}\}, & \text{if } a_2 \geq \bar{a}(z,\tilde{z}), \\ \zeta^a(z,a_2), & \text{if } a_2 < \bar{a}(z,\tilde{z}), \end{cases} \tag{13}$$

where we have defined

$$\zeta^a(z,a_2) \equiv \Big\{ \zeta : (1-\lambda)\left[u\left(z+\zeta\right) - u(z)\right] + \lambda\zeta = \psi_1 a_2 \Big\}. \tag{14}$$

Proof. The proof is straightforward, and it is, therefore, omitted. For a detailed proof of a similar bargaining problem, see Geromichalos and Herrenbrueck (2012). Below we provide an intuitive explanation of the bargaining solution. □

If $z + \tilde{z} \geq q^*$, the C-type should receive exactly as many real balances as she lacks in order to purchase q^* in the forthcoming LW market, i.e. $\zeta = q^* - z$. In contrast, if $z + \tilde{z} < q^*$, even if the two types pull together all their real balances, these will not allow the C-type to attain q^*. The second best, requires the N-type to give all her real balances to the C-type, $\zeta = \tilde{z}$. However, one

[11] Notice that one can re-arrange (10) to obtain $\psi_1\chi = \zeta + (1-\lambda)\left[u(z+\zeta) - u(z) - \zeta\right]$. This expression states that the real value of assets that the N-type receives as payment equals the value of real balances she is giving up, ζ, plus a fraction $1-\lambda$ (her bargaining power) of the surplus generated by the OTC transaction, i.e. the term $u(z+\zeta) - u(z) - \zeta$. Substituting for $\psi_1\chi$ from the last expression into (9), simplifies the bargaining problem to

$$\max_{\chi,\zeta} \lambda\left\{u\left(z+\zeta\right) - u\left(z\right) - \zeta\right\},$$

subject to (10) and the feasibility constraints $\chi \leq a_2$, $\zeta \leq \tilde{z}$. As is standard in proportional bargaining, the C-type's surplus equals a fraction λ of the total surplus generated by OTC trade.

should also ask whether the C-type has sufficient amounts of a_2 to compensate the N-type for the transfer of liquidity. This critical level of assets is defined in (11), and it depends on whether $z + \tilde{z}$ exceeds q^* or not. If $a_2 \geq \bar{a}(z, \tilde{z})$, the C-type is not constrained, and $\zeta = \min\{q^* - z, \tilde{z}\}$, as described above. In this case, the C-type gives up exactly $\bar{a}(z, \tilde{z})$ units of long term assets. When $a_2 < \bar{a}(z, \tilde{z})$, the C-type will not be able to purchase the desired amount of liquid assets, given by $\min\{q^* - z, \tilde{z}\}$. In that case, she will give away all her long maturity assets, $\chi = a_2$, and the transfer of real balances will be determined such that the sharing rule of the surplus between the two parties (equation (10)) is satisfied. Notice that the N-type's long term asset holdings do not affect the bargaining solution.

Having established the bargaining solutions in the OTC and LW markets, we now proceed to the derivation of the buyer's objective function and the description of her optimal behavior.

3.3 Objective Function and Optimal Behavior

In this sub-section, we characterize the optimal portfolio choice of the representative buyer. We will do so by deriving the buyer's objective function, i.e. a function that summarizes the buyer's cost and benefit from choosing any particular portfolio $(\hat{a}_1, \hat{a}_2)$. Substitute (7) and (8) into (6), and lead the resulting expression by one period to obtain

$$\begin{aligned}\mathbb{E}\left\{\Omega^i(\hat{a}_1, \hat{a}_2)\right\} = f\, V\left(\hat{a}_1 + \zeta, \hat{a}_2 - \chi\right) + (\ell - f)\, V\left(\hat{a}_1, \hat{a}_2\right) \\ + f W(\hat{a}_1 - \tilde{\zeta}, \hat{a}_2 + \tilde{\chi}) + (1 - \ell - f)\, W(\hat{a}_1, \hat{a}_2),\end{aligned} \tag{15}$$

where f is a shortcut for $f(\ell, 1 - \ell)$.

The four terms in (15) represent the benefit for a buyer who holds a portfolio $(\hat{a_1}, \hat{a_2})$ and turns out to be a matched C-type (with probability f), an unmatched C-type (with probability $\ell - f$), a matched N-type (with probability f), or an unmatched N-type (with probability $1 - \ell - f$), respectively. The expressions χ, ζ, and $\tilde{\chi}, \tilde{\zeta}$ are implicitly described by the solution to the OTC bargaining problem. In particular, we abuse notation slightly and define:

$$\chi = \chi(\hat{a}_1, \tilde{a}_1, \hat{a}_2), \quad \zeta = \zeta(\hat{a}_1, \tilde{a}_1, \hat{a}_2), \quad \tilde{\chi} = \chi(\tilde{a}_1, \hat{a}_1, \tilde{a}_2), \quad \tilde{\zeta} = \zeta(\tilde{a}_1, \hat{a}_1, \tilde{a}_2).$$

In these expressions, the first argument represents the C-type's real balances, the second argument represents the N-type's real balances, and the third argument stands for the C-type's long term asset holdings (recall from Lemma 2 that the N-type's long term asset holdings do not affect the bargaining solution). Terms with tildes stand for the representative buyer's beliefs about her potential counterparty's real balances and long term asset holdings in the OTC.[12]

[12] For instance, $\tilde{\zeta} = \zeta(\tilde{z}, \hat{z}, \tilde{a}_2)$ stands for the amount of real balances that the agent will give away if she is a matched N-type. This term depends on her own real balances ($\hat{z}$), and the real balances ($\tilde{z}$) and long term asset

Since each unit of asset that matures in the next period pays one unit of fruit before the LW market opens, it is understood that $\hat{z} = \hat{a}_1$, and from here on we describe optimal behavior in terms of the choice $(\hat{z}, \hat{a}_2)$ rather than $(\hat{a}_1, \hat{a}_2)$.

Next, we substitute W and V from (3) and (5), respectively, into (15). We insert the term $\mathbb{E}\{\Omega^i(\hat{z}, \hat{a}_2)\}$ into (2), and we focus on the terms inside the maximum operator of (2). We define the resulting expression as $J(\hat{z}, \hat{a}_2)$, and we refer to it as the buyer's objective function. After some manipulations, one can verify that

$$\begin{aligned} J(\hat{z}, \hat{a}_2) = & -\psi_1 \hat{z} - \psi_2 \hat{a}_2 \\ & + \beta\, f\, \left[u\left(\hat{z} + \zeta\right) + \hat{\psi}_1\left(\hat{a}_2 - \chi\right)\right] + \beta\, \left(\ell - f\right) \left[u\left(\hat{z}\right) + \hat{\psi}_1 \hat{a}_2\right] \\ & + \beta\, f\, \left[\hat{z} - \tilde{\zeta} + \hat{\psi}_1\left(\hat{a}_2 + \tilde{\chi}\right)\right] + \beta\, \left(1 - \ell - f\right) \left(\hat{z} + \hat{\psi}_1 \hat{a}_2\right). \end{aligned} \tag{16}$$

The two negative terms in the definition of J represent the cost of purchasing various amounts of the assets available in the economy.[13] The four positive terms in the definition of J admit similar interpretations as their counterparts in equation (15).

We can now proceed with the examination of the buyer's optimal choice of $(\hat{z}, \hat{a}_2)$. We will do so for any possible money and asset prices, and for any given beliefs about other agents' money and asset holdings. We know that the asset prices have to satisfy $\psi_1 \geq \beta$ and $\psi_2 \geq \beta\hat{\psi}_1$, since violation of these conditions would generate an infinite demand for the assets. The optimal behavior of the buyer is described formally in Lemma 3 below. Here, we provide an intuitive explanation of the buyer's optimal portfolio choice.

The objective function of the buyer depends on the terms $\chi, \zeta, \tilde{\chi}$, and $\tilde{\zeta}$, which, in turn, depend on the bargaining protocol in the OTC market. Given the buyer's beliefs $(\tilde{z}, \tilde{a}_2)$, she can end up in different branches of the bargaining solution, depending on her own choices of $(\hat{z}, \hat{a}_2)$. In general, the domain of the objective function can be divided into five regions in $(\hat{z}, \hat{a}_2)$-space, arising from three questions: (*i*) When the C-type and the N-type pool their real balances in the OTC market, can they achieve the first-best in the LW market? (*ii*) If I am a C-type, do I carry enough long term assets to compensate the N-type? (*iii*) If I am an N-type, do I expect a C-type to carry enough long term assets to compensate me? These regions are illustrated in Figure 2, and are described in detail as follows (for this discussion it is important to recall the definition of the asset cutoff term $\bar{a}(\,\cdot\,,\,\cdot\,)$ from Lemma 2).

1. $\hat{z} \in (q^* - \tilde{z}, q^*)$ and $\hat{a}_2 > \bar{a}(\hat{z}, \tilde{z})$.

holdings ($\tilde{a}_2$) of her trading partner (a C-type). The terms χ, ζ, and $\tilde{\chi}$ admit similar interpretations.

[13] In the objective function, the term $-\psi_1 \hat{z}$ appears as the cost of purchasing assets that mature in the next period. However, we know that the term $\psi_1 a_2$ is also present in the agent's value function (see equation (2)), so that, practically, the cost of leaving the CM with $\hat{z}$ units of assets that mature tomorrow is $-\psi_1(\hat{z} - a_2)$. However, the term $\psi_1 a_2$ only has a level effect, and it does not change the optimal choice of $\hat{z}$, with the exception that any choice of the agent should respect the restriction $\hat{z} - a_2 \geq 0$.

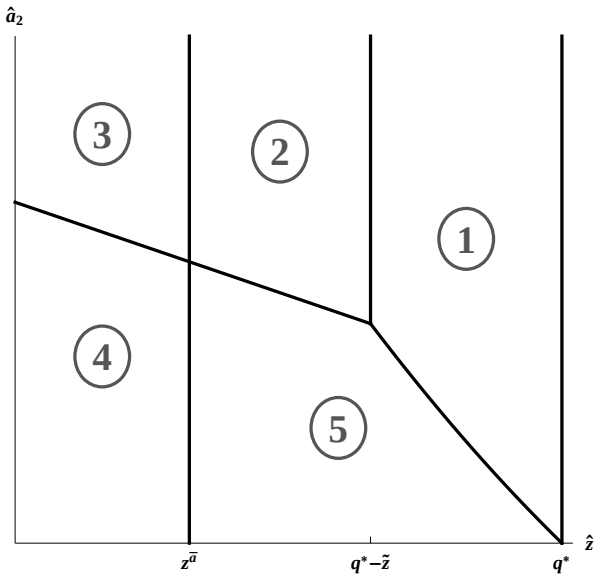

Figure 2: Regions of the individual choice problem.

In this region, the real balance holdings of the C-type and the N-type together allow the C-type to purchase q^* in the LW market. If the agent is a C-type, her long term asset holdings are enough to compensate an N-type for her real balances. If the agent is an N-type, the potential counterparty may or may not carry enough long term assets to purchase the first-best level of real balances, $q^* - \tilde{z}$, but that is a level effect on $G(\hat{z}, \hat{a}_2)$ and does not affect the optimal choice.[14]

2. $\hat{z} < q^* - \tilde{z}$, $\hat{a}_2 > \bar{a}(\hat{z}, \tilde{z})$, but $\tilde{a}_2 < \bar{a}(\tilde{z}, \hat{z})$.
   Here there are not enough real balances in an OTC match to allow the C-type to purchase q^* in the LW market. If a C-type, the agent carries enough long term assets to buy all the real balances of the N-type, but if an N-type, the agent does not expect the C-type counterparty to carry enough long term assets to buy all of the agent's real balances.

3. $\hat{z} < q^* - \tilde{z}$, $\hat{a}_2 > \bar{a}(\hat{z}, \tilde{z})$, and $\tilde{a}_2 > \bar{a}(\tilde{z}, \hat{z})$.
   There are not enough real balances in an OTC match to allow the C-type to purchase q^* in the LW market. In an OTC match, the agent expects all of the real balances of the N-type to be traded for less than all of the long term assets of the C-type (regardless of whether the buyer in question is the C or the N-type).

4. $\hat{z} < q^* - \tilde{z}$, $\hat{a}_2 < \bar{a}(\hat{z}, \tilde{z})$, but $\tilde{a}_2 > \bar{a}(\tilde{z}, \hat{z})$.
   There are not enough real balances in an OTC match to allow the C-type to purchase q^* in the LW market. If a C-type, the agent does not carry enough long term assets to buy all the

[14] Since here the objective is to describe the buyer's optimal behavior, we focus on how different choices of $(\hat{z}, \hat{a}_2)$ lead to different branches of the OTC bargaining protocol. In Region 1, the buyer is not certain whether her C-type counterparty is asset constrained or not, but she also does not care. What determines Region 1 is that, conditional on being an N-type, the buyer's real balances never affect the terms of trade.

real balances of the N-type, but if an N-type, the agent expects the C-type counterparty to carry enough long term assets to buy all of her real balances.

5. $\hat{a}_2 < \bar{a}(\hat{z}, \tilde{z})$, and either $\tilde{a}_2 < \bar{a}(\tilde{z}, \hat{z})$ or $\hat{z} \in (q^* - \tilde{z}, q^*)$.
   If a C-type, the agent does not carry enough long term assets to buy all the real balances of the N-type. If an N-type, the agent expects not to give away all of her real balances, either because the C-type counterparty does not carry enough long term assets to afford it, or because she does not need all of those real balances. This distinction does not affect the buyer's optimal choice.

We can now state the most important facts about the optimal choice of the representative buyer:

Lemma 3. *Taking prices, $(\psi_1, \hat{\psi}_1, \psi_2)$, and beliefs, $(\tilde{z}, \tilde{a})$, as given, and assuming that $\psi_1, \hat{\psi}_1 \geq \beta$ and $\psi_2 \geq \hat{\psi}_1$, then the optimal choice of the representative agent, $(\hat{z}, \hat{a}_2)$, satisfies:*

a) If the optimal choice $(\hat{z}, \hat{a}_2)$ is strictly within any region, or on the boundary of Region 1 with any other region, it satisfies the first-order condition $\nabla J = \mathbf{0}$.

b) If $\psi_1 > \beta$ and $\psi_2 = \beta\hat{\psi}_1$, the optimal $\hat{z}$ is unique, and any $\hat{a}_2$ is optimal as long as $(\hat{z}, \hat{a})$ is in Regions 1, 2, or 3 (or on their boundaries).

c) If $\psi_1 > \beta$ and $\psi_2 > \beta\hat{\psi}_1$, the optimal choice is unique, and it lies in Regions 4 or 5 or on their boundaries with Regions 2 and 3.

Moreover, let $J^i(\hat{z}, \hat{a}_2)$, $i = 1, ..., 5$, denote the objective function in Region i, and $J^i_k(\hat{z}, \hat{a}_2)$, $k = 1, 2$, its derivative with respect to the k-th argument. Then, we have:

$$\beta^{-1} J^1_1(\hat{z}, \hat{a}_2) = -\frac{\psi_1}{\beta} + 1 + (\ell - \lambda f)\left[u'(\hat{z}) - 1\right], \tag{17}$$

$$\beta^{-1} J^2_1(\hat{z}, \hat{a}_2) = -\frac{\psi_1}{\beta} + 1 + (\ell - \lambda f)\left[u'(\hat{z}) - 1\right] + \lambda f\left[u'(\hat{z} + \tilde{z}) - 1\right], \tag{18}$$

$$\beta^{-1} J^3_1(\hat{z}, \hat{a}_2) = -\frac{\psi_1}{\beta} + 1 + (\ell - \lambda f)\left[u'(\hat{z}) - 1\right] + f\left[u'(\hat{z} + \tilde{z}) - 1\right], \tag{19}$$

$$\begin{aligned}\beta^{-1} J^4_1(\hat{z}, \hat{a}_2) = -\frac{\psi_1}{\beta} + 1 + \ell\left[u'(\hat{z}) - 1\right] + (1 - \lambda) f\left[u'(\hat{z} + \tilde{z}) - 1\right] + \dots \\ \dots + \lambda f \frac{u'\left[\hat{z} + \zeta^a(\hat{z}, \hat{a}_2)\right] - u'(\hat{z})}{(1 - \lambda) u'\left[\hat{z} + \zeta^a(\hat{z}, \hat{a}_2)\right] + \lambda},\end{aligned} \tag{20}$$

$$\beta^{-1} J^5_1(\hat{z}, \hat{a}_2) = -\frac{\psi_1}{\beta} + 1 + \ell\left[u'(\hat{z}) - 1\right] + \lambda f \frac{u'\left[\hat{z} + \zeta^a(\hat{z}, \hat{a}_2)\right] - u'(\hat{z})}{(1 - \lambda) u'\left[\hat{z} + \zeta^a(\hat{z}, \hat{a}_2)\right] + \lambda}, \tag{21}$$

$$J^1_2(\hat{z}, \hat{a}_2) = J^2_2(\hat{z}, \hat{a}_2) = J^3_2(\hat{z}, \hat{a}_2) = -\psi_2 + \beta\hat{\psi}_1, \tag{22}$$

$$J^4_2(\hat{z}, \hat{a}_2) = J^5_2(\hat{z}, \hat{a}_2) = -\psi_2 + \beta\hat{\psi}_1 \left\{1 - f + f \frac{u'\left[\hat{z} + \zeta^a(\hat{z}, \hat{a}_2)\right]}{(1 - \lambda) u'\left[\hat{z} + \zeta^a(\hat{z}, \hat{a}_2)\right] + \lambda}\right\}, \tag{23}$$

where $\zeta^a(\,\cdot\,,\,\cdot\,)$ was defined in (14).

Proof. See Appendix A. □

Lemma 3 formally describes the optimal behavior of the representative buyer. Given the results stated in the lemma, one can describe in detail the demand functions for the various assets. Although interesting, this analysis is not essential for understanding the main results of the paper, hence, it is relegated to the web appendix. We are now ready to discuss equilibrium.

3.4 Definition of Equilibrium and Preliminary Results

We restrict attention to symmetric steady-state equilibria, where all agents choose the same portfolios, and the real variables of the model remain constant over time. Before stating the definition of a steady-state equilibrium, we will examine which regions of the individual choice problem (in Figure 2) can be reached in such an equilibrium. Aggregate real balances Z are equal to the aggregate supply of maturing short term assets (A_1), and the aggregate supply of long term assets is exogenously given (A_2).

First, symmetry rules out Regions 2 and 4, since a C-type and an N-type buyer are ex ante identical. Second, recall that by definition, A_1 refers to the total stock of assets which mature one period from now, including both newly issued short term assets and previously issued long term assets. So the mere assumption of a steady state constrains A_1 to be at least as large as A_2. This is sometimes called the "cascade effect", and combinations of $A_1 < A_2$ are shaded gray in the right panel of Figure 3. Third, recall our constraint that buyers cannot sell off-the-run short term assets in the CM; at most, they can refrain from buying newly issued short term assets. If this constraint binds on anyone, their value function is no longer linear and the model will become intractable. We therefore restrict attention to equilibria in which the post-CM holdings of short term assets ($\hat{a}_1$ for an individual, equal to A_1 in symmetric equilibrium) exceed the pre-CM holdings (a_2 for an individual) for everyone. In symmetric equilibrium, buyers who were C-types (asset sellers) in the preceding OTC market enter the CM with $a_2 = A_2 - \chi(A_1, A_1, A_2)$, and buyers who were N-types (asset buyers) in the preceding OTC market hold $A_2 + \chi(A_1, A_1, A_2)$. Combinations of (A_1, A_2) which violate the $A_1 \geq A_2 + \chi(A_1, A_1, A_2)$ constraint are shaded pink in the right panel of Figure 3 and labeled "inadmissible".

It is important to emphasize that all we do here is rule out a subset of parameter values where we cannot both characterize equilibria and also keep the constraint $\hat{a}_1 - a_2 \geq 0$ in the buyer's CM problem. In those equilibria which we do characterize, the constraint never binds. Alternatively, we could eliminate the constraint and allow buyers to sell off-the-run short term assets in the CM. But if we did, none of the technicalities of the paper (demand functions with kinks, rich asset prices, etc.) would change; the only change is that we would not have to rule out the "inadmissible" parameter region, a benefit which we think is relatively small.

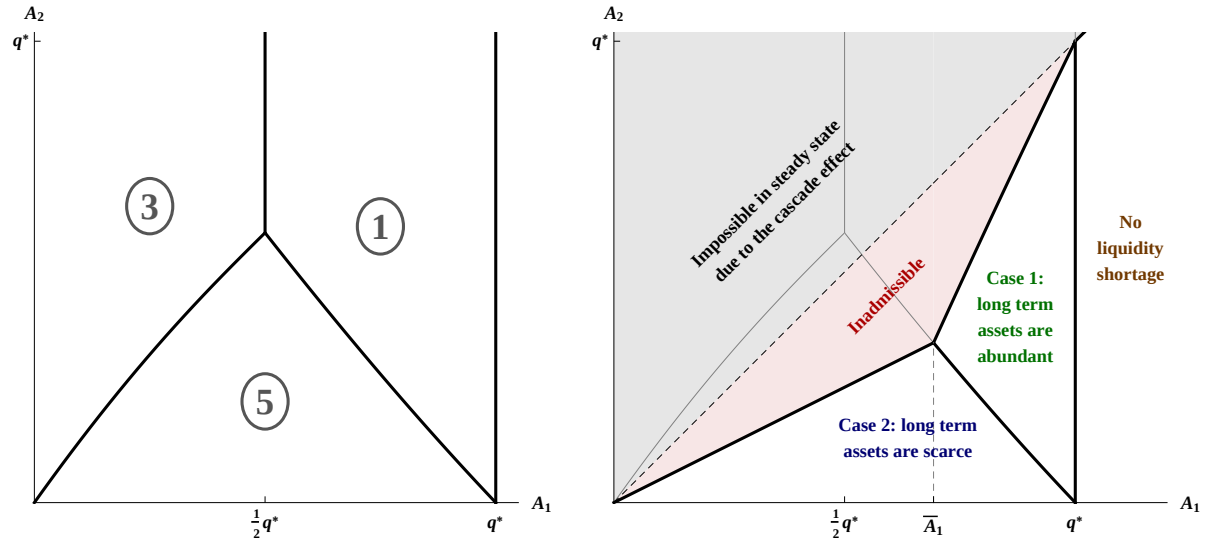

Figure 3: Aggregate regions of equilibrium.

The surviving regions of aggregate equilibrium are depicted in the left panel of Figure 3, and we will refer to them as the "aggregate regions", as opposed to the "individual regions" depicted in Figure 2. The aforementioned constraints rule out combinations of low A_1 and high A_2. Therefore, they imply that equilibrium will never be in Region 3, where the real balances of the N-type restrict OTC trade and equilibrium becomes harder to analyze, as long as the following restriction on structural parameters is satisfied:[15]

$$\frac{1+(1-\lambda)\left[\frac{u(q^*)-u(q^*/2)}{q^*/2}-1\right]}{1+(\ell-\lambda f)\left[u'(q^*/2)-1\right]}>\frac{\beta}{2}. \tag{24}$$

Henceforth, we assume that the parameters of the model satisfy the inequality stated in (24).

With the above constraints satisfied, and ignoring for now the trivial case $A_1 \geq q^*$, only two cases remain on aggregate (depicted in the right panel of Figure 3):

Case 1. Agents carry enough real balances and long term assets so that, when matched in the OTC market, the C-type can acquire sufficient liquidity in order to achieve the first-best in the LW market.

Case 2. Agents carry so few long term assets that, when matched in the OTC, the C-type will sell all of her long term assets but not obtain enough of the N-type's real balances in order to achieve the first-best in the LW market.

[15] This restriction is derived from the condition that the line $A_1 = 2A_2$ should pass below the meeting point of Regions 1, 3, and 5 in Figure 3. While it is possible to construct a counterexample, the restriction is satisfied for a wide range of utility functions if f is close to ℓ (C-types have a high probability of matching). Let us also point out that this restriction simplifies the analysis significantly without ruling out any interesting results. In particular, Results 1-3 stated in Section 3.5 would still go through without this additional assumption. See Geromichalos and Herrenbrueck (2012) for an analysis of Region 3 equilibria in a model with money and a single real asset.

Case 1 represents the admissible part of Region 1, the region of abundance of the long maturity asset, and Case 2 represents the admissible part of Region 5, the region where this asset is scarce in OTC trade.

Definition 1. A symmetric steady-state equilibrium is a list $\{\psi_1, \psi_2, \chi, \zeta, q_1, q_2\}$, where q_1 denotes the amount of goods exchanged in the LW market when the buyer was not matched in the preceding OTC market, and the term q_2 is the amount of goods exchanged in the LW market when the buyer *was* matched. The equilibrium objects satisfy:

i. The representative buyer behaves optimally under the equilibrium prices ψ_1, ψ_2.

ii. We have $q_1 = A_1$ and either $q_2 = q^*$ (if equilibrium lies in Region 1) or $q_2 = \tilde{q}(A_1, A_2)$ (if equilibrium lies in Region 5), where $\tilde{q}$ solves $(1-\lambda)\left[u(\tilde{q}) - u(A_1)\right] + \lambda\left(\tilde{q} - A_1\right) = \psi_1 A_2$.

iii. The terms (χ, ζ) satisfy (12) and (13) evaluated at the aggregate quantities A_1 and A_2.

iv. Markets clear at symmetric choices, and expectations are rational: $\hat{z} = \tilde{z} = A_1$, and $\hat{a}_2 = \tilde{a}_2 = A_2$.

Lemma 4. *If $A_2 + \chi[A_1, A_1, A_2] \leq A_1 \leq q^*$ are satisfied, then a symmetric steady-state equilibrium exists and is unique.*

Proof. See Appendix A. □

Having formally described the definition of a steady-state equilibrium and guaranteed its existence and uniqueness, the next task is to characterize such equilibria, and to describe the equilibrium variables as functions of the exogenous supply parameters A_1 and A_2. But before we begin, we define the cutoff levels of short term and long term asset supply that will separate the classes of equilibria. If A_1 is small, then long term assets must be scarce in OTC trade because any admissible level of A_2 will be small, too. If A_1 is large, long term assets may either be abundant or scarce. The relevant cutoff for A_1, indicated in Figure 3, is given by:

$$\bar{A}_1 \equiv \left\{ A_1 : \frac{1}{2}A_1 = \frac{(1-\lambda)\left[u(q^*) - u(A_1)\right] + \lambda(q^* - A_1)}{\beta + \beta(\ell - \lambda f)\left[u'(A_1) - 1\right]} \right\}.$$

Condition (24) is equivalent to $\bar{A}_1 > q^*/2$. Next, we define the cutoff level of long term asset supply as a function of short term asset supply that describes the upper boundary of the region of admissible equilibria where long term assets are scarce:

$$\bar{A}_2(A_1) \equiv \min\left\{ \frac{1}{2}A_1, \frac{(1-\lambda)\left[u(q^*) - u(A_1)\right] + \lambda(q^* - A_1)}{\beta + \beta(\ell - \lambda f)\left[u'(A_1) - 1\right]} \right\}.$$

3.5 Characterization of Equilibrium

We begin this section with an intuitive description of the results presented in Propositions 1 and 2. If the supply of maturing assets (A_1) is plentiful, short term assets alone are enough to satisfy the liquidity needs of the economy (for trade in the LW market), and there is no role for OTC trade. On the other hand, if A_1 is insufficient to satisfy the liquidity needs of the economy (which we consider the interesting case), this has two important implications for asset prices. First, ψ_1 will carry a *liquidity premium* (i.e. $\psi_1 > \beta$), because the marginal unit of short term assets is not only a good store of value, but it can also increase consumption in the LW market. Second, trade in the OTC market becomes valuable as it helps agents to efficiently reallocate real balances after the need for liquidity has been revealed. Consequently, the long term assets can potentially also carry a liquidity premium, not because they can facilitate trade in the LW market directly, but because they can be used in the OTC market in order to purchase liquid assets. In particular, ψ_2 will include a liquidity premium if the supply A_2 is relatively scarce.

We now describe these results in a formal way.

Proposition 1. *If $A_1 \geq q^*$, no trade occurs in the OTC market, and asset prices always equal their fundamentals: $\psi_i = \beta^i$ for $i = 1, 2$.*

Proof. See Appendix A. □

When $A_1 \geq q^*$, the supply of maturing short term assets suffices to cover the liquidity needs of the economy (i.e. the need for trade in the anonymous LW market). This has the following consequences. First, since agents already bring with them sufficient liquidity in order to purchase q^*, there is no role for trade in the OTC market. Second, since assets are issued in a competitive market, ψ_1 will reflect the benefit of holding one additional unit of short term assets. But since here $A_1 \geq q^*$, the marginal unit of short term assets is only good as a store of value, and not as a facilitator of trade in the LW market (a role which this asset now serves inframarginally). Thus, the unique equilibrium price must be $\psi_1 = \beta$. Finally, with no trade in the OTC market, long term assets cannot be valued for any (direct or indirect) liquidity properties either, which means that $\psi_2 = \beta^2$.

Henceforth, we maintain the assumption $A_1 < q^*$, which we consider the interesting case (since $A_1 \geq q^*$ implies no trade in the OTC market). Since the focus of the paper is on asset prices (and, consequently, asset yields), we relegate the analysis of equilibrium production in the LW market to Appendix B.

Proposition 2. *The equilibrium prices of assets depend on the values of A_1 and A_2. We have two cases:*

Case 1: *If $A_2 \geq \bar{A}_2(A_1)$, which is only admissible if also $A_1 > \bar{A}_1$, then equilibrium is in Region 1, and:*

$$\psi_1 = \beta(1+\theta),$$
$$\psi_2 = \beta^2(1+\theta).$$

Case 2: *If $A_2 < \bar{A}_2(A_1)$, then equilibrium is in Region 5, and:*

$$\psi_1 = \beta(1+\theta+\rho),$$
$$\psi_2 = \beta^2(1+\rho)(1+\theta+\rho).$$

The terms $\theta \geq 0$ and $\rho \geq 0$ denote liquidity premia and are explained below.

Proof. See Appendix A. □

The price of short term assets will include a *direct* liquidity premium because these assets play a monetary role in this economy: they represent claims to the only possible medium of exchange (dividend) in the LW market, and the claims always pay off in time to take advantage of an opportunity to trade in the LW market. The term θ measures this direct liquidity premium. In terms of the equilibrium objects q_1 and q_2 (from Definition 1), it is defined by:

$$\theta = \left[\ell - \frac{\lambda f}{(1-\lambda)u'(q_2)+\lambda}\right][u'(q_1)-1]. \tag{25}$$

The price of long term assets can include a liquidity premium for two reasons. First, because long term assets will become short term assets in the next period; hence, the term θ appears in the equation for ψ_2, but only through agents' expectations of ψ_1. Second, because long term assets can be used in the OTC market in order to purchase liquid assets; the assets that do not mature today have *indirect* liquidity properties because they help agents bypass the cost of holding liquid assets, θ (which is positive when $A_1 < q^*$). The term ρ measures this indirect liquidity premium, and it is defined by:

$$\rho = \frac{\lambda f[u'(q_2)-1]}{(1-\lambda)u'(q_2)+\lambda}. \tag{26}$$

The reason why the terms q_2 (through θ) and ρ also appear in the price of the short term asset in Case 2 is that OTC rebalancing is costly for the C-type when $\lambda < 1$, and carrying more short term assets helps C-types avoid this cost.[16]

The results reported in Proposition 2 are illustrated in Figure 4. The left panel reproduces

[16] Note that the equations in Case 1 are a special case of those in Case 2 because $q_2 = q^*$ in Case 1. Since $u'(q^*) = 1$ we then have $\rho = 0$, and while θ is still positive, its definition becomes simpler: $\theta = (\ell - \lambda f)[u'(q_1) - 1]$. In Case 2, we have $q_2 = \tilde{q}(A_1, A_2)$ (see Definition 1), which is increasing at a rate less than 1 in both A_1 and A_2, and converges to q^* as $A_2 \to \bar{A}_2(A_1)$.

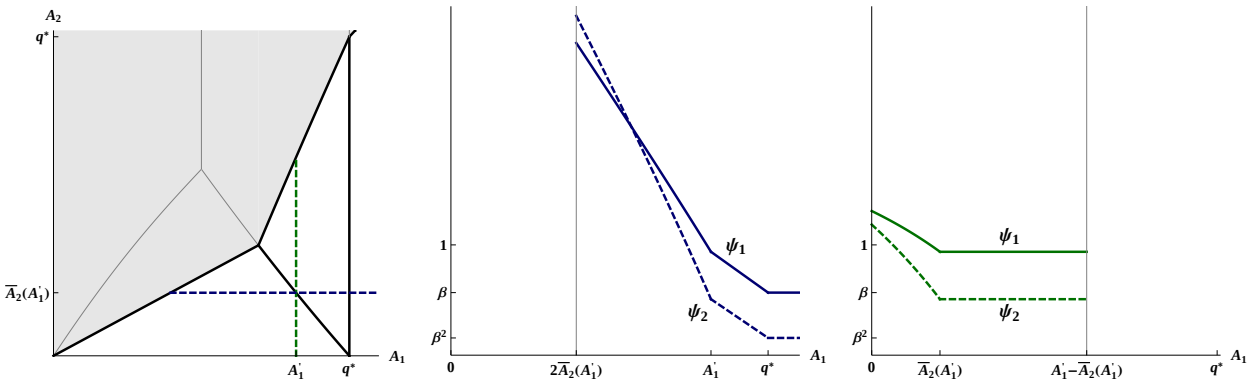

Figure 4: Left panel: location of a fixed $A_1' \in (\bar{A}_1, q^*)$. Middle panel: equilibrium asset prices as functions of A_1 for $A_2 = \bar{A}_2(A_1')$. Right panel: ditto as functions of A_2 for fixed $A_1 = A_1'$.

the regions of equilibrium and fixes two particular levels $A_1' > \bar{A}_1$ and $A_2' \equiv \bar{A}_2(A_1')$. The middle panel depicts both asset prices as a function of A_1 for $A_2 = A_2'$, and the right panel does the same as a function of A_2 for $A_1 = A_1'$. Both prices are (weakly) decreasing as a function of the supply of either asset, indicating that the opportunity of OTC rebalancing makes assets of different maturities imperfect substitutes. The mathematical reason for this effect is the following. As long as $A_1 < q^*$ (so that liquidity is valued), the direct liquidity premium θ is positive. In Case 1, the indirect liquidity premium ρ is zero and θ is decreasing in A_1 but independent of A_2. In Case 2, both liquidity premia are positive and strictly decreasing in A_1 and A_2.

3.6 Empirically Supported Predictions of the Model

3.6.1 The Term Premium of Long Term Assets

In order to discuss the model's predictions for the term structure of interest rates, we first define the term premium between long and short term assets in terms of the return differential:

$$\tau_{1,2} \equiv r_2 - r_1,$$

where r_i, $i = 1, 2$, was defined by equation (1). Note that, given the restriction to steady-state equilibria (so that r_i is constant), in the absence of any liquidity considerations, the expectations hypothesis would imply a flat yield curve. However, the model predicts the existence of a positively-sloped yield curve, which is a well-established feature of bond yields.

Result 1. *Assume that $A_1 < q^*$ and that the conditions for existence of equilibrium (Lemma 4) are satisfied. Then the term premium is strictly positive; formally, $\tau_{1,2} > 0$.*

Result 1 reveals that our model delivers a positive term premium, as long as the supply of

long term maturities is relatively scarce.[17] To see why this result is true, consider first Case 1 in which long term assets are relatively plentiful. Here, we have:

$$\tau_{1,2} = \frac{1}{\beta}\left[(1+\theta)^{-0.5} - (1+\theta)^{-1}\right],$$

which is positive since $A_1 < q^*$ implies a positive direct liquidity premium $\theta > 0$ (Proposition 2). If equilibrium lies in the regions of scarcity of long term assets (Case 2), we positive direct and indirect liquidity premia ($\theta > 0$ and $\rho > 0$). In this case:

$$\tau_{1,2} = \frac{1}{\beta}\left([(1+\rho)(1+\theta+\rho)]^{-0.5} - (1+\theta+\rho)^{-1}\right).$$

Again, given that in Case 2 we have $A_2 < \bar{A}_2(A_1)$, we conclude that $\theta > 0$ and $\tau_{1,2} > 0$.

The analysis above indicates that the positive term premium is crucially linked to the existence of liquidity premia in asset prices. Short term assets are a way to obtain a medium of exchange at exactly the moment when needed, and, under the assumption that $A_1 < q^*$, they always carry a liquidity premium that reflects the ability of the marginal unit to increase LW consumption. On the other hand, long term assets do not have this property, so that agents who hold long maturities must be compensated for their relative illiquidity in the form of a positive term premium.

The term premium $\tau_{1,2}$ is positive even if the long term assets carry indirect liquidity properties (Case 2) due to their ability to help agents acquire liquid assets in the OTC market. But say we want to examine which parameters would drive the term premium close to zero. Clearly, this requires θ, which measures the liquidity advantage of short term over long term assets, to go to zero. Using its definition in equation (25), we see that there are three ways this can happen: (a) if $\ell = 0$, so that there is no demand for liquidity; (b) if $A_1 \geq q^*$, so that the demand for liquidity is fully satisfied; or (c) if the multiplier of the term $u'(A_1) - 1$ in equals zero. This multiplier is given by $\ell - \lambda f[(1-\lambda)u'(q_2) + \lambda]^{-1}$, and it will be equal to zero only if $\ell = f$ (C-types match with probability 1) *and also* $\lambda = 1$ (C-types have all the bargaining power). This result is very intuitive. A buyer will be willing to hold long term assets at yield $r_2 = r_1$ only if they are as liquid as short term assets, and this will be true only if the C-type (the type of agent who needs liquidity) is guaranteed to match in the OTC market and is able to extract the whole surplus of that match.

The effect of asset supply on the term premium is complicated to fully describe using the definition $\tau_{1,2}$. But the ultimate source of a positive term premium in our model is the parameter θ, which is strictly decreasing in A_1 (both directly through q_1 and indirectly through q_2), strictly decreasing in A_2 in Case 2, and unaffected by A_2 in Case 1. Furthermore, if θ and ρ are

[17] If $A_1 \geq q^*$, we know from Proposition 1 that assets will always be priced at their fundamental value, i.e. $\psi_i = \beta^i$, for $i = 1, 2$. This, in turn, implies (by (1)) that $r_i = 1/\beta - 1$, for $i = 1, 2$, so that $\tau_{1,2} = 0$.

small, we can approximate the term premium with $\tau_{1,2} \approx \theta/(2\beta)$, which will then also decrease in the supply of either asset.

At this point, two observations are in order. First, the existence of the OTC market is not a necessary condition for Result 1. In fact, if one shuts down the OTC market, not only will a positive term premium survive, but it will actually attain its maximum value. This is simply true because shutting down the OTC is a special case of our model where $f = 0$, which is equivalent to maximizing the OTC market frictions. Second, in our model short term assets mature *just in time* to take advantage of consumption opportunities, i.e. right before the LW market opens. This feature is an artifact of our discrete time model and our timing specification, and it is also not a necessary condition for Result 1. What is essential for Result 1 is that short term assets allow agents to bypass the costs associated with liquidating long term assets in the frictional OTC market. Therefore, Result 1 will hold true in any extension of the model where short term assets mature *closer* to (rather than right before) possible liquidity needs of the agent.[18]

3.6.2 The Effect of Secondary Market Liquidity on Asset Returns

One of the key insights of our model is that the issue price of long maturity assets is crucially affected by the liquidity of the secondary asset market, i.e. how easy it is for agents to liquidate these long maturity assets. To highlight the importance of this liquidity mechanism for equilibrium asset returns, we conduct the following experiment: we extend the baseline model (with $N = 2$) to include a second set of assets whose only difference from the original assets studied in previous sections is that they *cannot be traded* in secondary markets (the new assets are present *only* in Section 3.6.2). In any other aspect, the new assets are identical to the original ones. The new assets come in fixed supplies denoted by B_i, $i = 1, 2$. Agents can purchase assets of maturity $i = 1, 2$ at the ongoing market price p_i (in the CM), and each unit of asset of maturity $i = 1, 2$ purchased in period t delivers one unit of (the same) fruit before the LW market of period $t + i$ opens. We will refer to the new assets as CDs (i.e. certificates of deposit), since a unique characteristic of these assets is that they have to be held to maturity.

As long as the supply of short term assets is not so large as to fully satisfy the liquidity needs of the economy, the issue price of long maturities will be higher for the assets that can be traded in secondary markets, reflecting the indirect liquidity premium.

Result 2. *Suppose that $A_1 + B_1 < q^*$, and the parameters are such that equilibrium lies in Region 5. Then, $\psi_1 = p_1 = (1 + \theta + \rho)$, $\psi_2 = \beta(1 + \rho)\psi_1$, $p_2 = \beta p_1$, and $\rho > 0$, so that $\psi_2 > p_2$. Moreover, the*

[18] For instance, one can envision a continuous time version of the model, where consumption opportunities arrive randomly, and they can wait for a period equal to $\tau > 0$. An agent who has an opportunity at time t to purchase assets that pay 1 dollar either in $t + 3$ or in $t + 6$ (say, months) can use the short term assets (and only these) to fund any consumption opportunity that may arise in the interval $[t+3-\tau, t+3]$. If, however, these assets are not enough, the agent will need to liquidate long term assets in the frictional OTC market. In this environment, the agent will still demand a premium in order to purchase the relatively more illiquid long term assets.

indirect liquidity premium ρ is decreasing in A_2.

Result 2 is a straightforward generalization of Proposition 2. Short term assets of both types are perfect substitutes (as both pay off in time to use the fruit as medium of exchange), so the existence of "interesting equilibria" (i.e. equilibria with liquidity premia) requires $A_1 + B_1 < q^*$. If this condition is satisfied, ψ_1 and p_1 will include the usual liquidity premium, and they will be equal. The price p_2 will include a liquidity premium only because long term CDs will become short term CDs in the next period, i.e. $p_2 = \beta p_1$. In contrast, ψ_2 can include an additional indirect liquidity premium, indicated by ρ, which reflects the assets' property to help agents avoid the cost of holding liquid assets. Thus, if A_2 is relatively scarce, we have $\psi_2 > p_2$.[19]

In terms of asset yields (rather than prices), letting r_2^{CD} denote the interest rate on long term CDs, Result 2 indicates that $r_2^{CD} - r_2^A > 0$. Moreover,

$$r_2^{CD} - r_2^A = \frac{1}{\beta\psi_1}\left[1 - (1+\rho)^{-0.5}\right] \approx \frac{\rho}{2\beta\psi_1}.$$

Within the region of relative scarcity of A_2 (the analogue of Case 2 in the baseline model), this expression is decreasing in A_2, because the indirect liquidity premium ρ is decreasing in A_2: a scarce A_2 makes the service that long term assets provide (helping agents avoid the holding cost of liquid assets) more valuable (Proposition 2).

Krishnamurthy and Vissing-Jorgensen (2012) provide direct evidence in support of these findings. The authors compare the yields on 6-month FDIC-insured CDs and 6-month treasury bills over the 1984-2008 period. Both assets are default-free, but, as we already mentioned, CDs have to be held to maturity, which is not the case for T-bills. Consequently, the authors suggest, any spread reflects the higher liquidity of T-bills. They report that, over the sample period, the spread was 2.3 percentage points on average, and was negatively related to the supply of T-bills. As Result 2 reveals, the model is consistent with both of these findings.

More generally, our model predicts that, *ceteris paribus*, equilibrium prices (yields) are increasing (decreasing) in the ease with which agents can trade assets in the secondary OTC market (or, more formally, $\partial\psi_2/\partial f > 0$ or $\partial r_2/\partial f < 0$, either of which follows immediately from Proposition 2). This finding is consistent with Gürkaynak, Sack, and Wright's (2010) analysis of the yield curve for inflation-indexed Treasury debt (i.e. TIPS). In particular, the authors demonstrate that, over the period from 1999 to 2005, the TIPS yields have, in general, fallen as market liquidity (measured by trading volume) in the TIPS market has increased.

A direct consequence of Result 2 is that the yield curve is steeper for assets that trade in less liquid secondary markets: we expect long term yields to reflect a liquidity differential, as above, but short term yields less so because these assets can be held to maturity. This finding is also

[19] In fact, as long term CDs cannot be traded before maturity at all, their term premium is as large as it can be given the other parameters of the economy, and always at least as large as the term premium for the A-assets.

empirically supported. For example, the yield curve for municipal bonds, which are known for trading in fairly illiquid secondary markets, is especially steep; see Green (1993).

3.6.3 The On-the-run Phenomenon

One interesting feature of our model is that N-type agents who, in the OTC market of period t, purchase assets issued at $t-1$ and maturing at $t+1$, could also obtain identical assets (maturing at $t+1$) in the forthcoming CM (of period t). Therefore, our model provides a framework in which one can compare the price of on-the-run short term assets with the price of older assets (off-the-run) which mature on the same date. Warga (1992) documents that the return of an off-the-run portfolio exceeds, on average, the return of an on-the-run portfolio with similar duration. Our model is consistent with this observation.

Result 3. *Assume that $A_1 < q^*$ and that the conditions for existence of equilibrium (Lemma 4) are satisfied. Define the (real) price of two-period assets (issued in the previous period and maturing in the next one) in the OTC market, $\psi_o \equiv \zeta/\chi$, where ζ, χ represent equilibrium objects (Definition 1). Comparing ψ_o with the issue price of assets that mature in the next period, ψ_1, we obtain:*

$$\psi_1 = \psi_o \left[(1-\lambda) \frac{u(q_2) - u(q_1)}{q_2 - q_1} + \lambda \right]. \tag{27}$$

In any equilibrium, $\psi_o < \psi_1$.

The term $[u(q_2) - u(q_1)]/[q_2 - q_1]$ represents the average surplus created by one unit of real balances traded in the OTC market. Since u is strictly concave, this average surplus must exceed the marginal surplus created by the last unit of real balances. Therefore,

$$\frac{u(q_2) - u(q_1)}{q_2 - q_1} > u'(q_2) \geq u'(q^*) = 1,$$

which establishes $\psi_o < \psi_1$.

The assets that are sold by N-types in period t's OTC market (issued at $t-1$ and maturing at $t+1$) have the same maturity structure as the short term assets issued in period t's CM. Hence, one might expect that their prices should be equal. This argument fails to recognize two important facts. First, the seller of off-the-run assets is not the same agent as the seller (issuer) of on-the-run assets.[20] Second, the very structure of the markets in which these two types of assets are traded is different. With respect to the first point, a seller of off-the-run assets is an agent who received a consumption opportunity (a C-type) and who, typically, is short of liquidity.

[20] In fact, here we remain agnostic as to who is the issuer of these assets by treating them as "Lucas trees". Geromichalos and Herrenbrueck (2015) study the problem of an issuer of liquid assets in a related framework.

This agent will be desperate for the N-type's liquidity and more willing to sell assets at a low price. Moreover, ψ_o is determined in an OTC market characterized by search and bargaining. Hence, while ψ_1 reflects the fundamental properties of short term assets (the marginal benefit of holding one extra unit), ψ_o represents the terms of trade that implement the "correct" sharing rule of the surplus generated during OTC trade. As long as the N-type has some bargaining power ($\lambda < 1$), she will always extract a fraction of the surplus and purchase assets at price $\psi_o < \psi_1$. This point becomes clear by noticing that $\psi_o = \psi_1$ only if $\lambda = 1$ (equation (27)).

Vayanos and Weill (2008) also provide a theoretical explanation of the on-the-run phenomenon. They build a model where on-the-run bonds are more valuable because they are more liquid than their off-the-run counterparts and because they constitute better collateral for borrowing in the repo market (a phenomenon known as "specialness").[21] Importantly, in their model, both of these advantages of on-the-run assets arise endogenously and simultaneously.

It should be pointed out that, in the data, on-the-run bonds command a premium even in secondary (i.e. OTC) trading. The model of Vayanos and Weill (2008) is able to capture this regularity, while our paper cannot, since, by construction, on-the-run bonds are only traded in the Walrasian market. In general, the model presented here was not built with the intention to capture the on-the-run phenomenon. However, we think that our model highlights a new, interesting, and extremely intuitive channel through which the on-the-run phenomenon could be explained. In particular, our model suggests that sellers of off-the-run bonds are agents who are in greater need of liquidity, and, hence, more eager to sell assets at a lower price.

4 Extensions of the Baseline Model

4.1 Equilibrium in the Model with Money

Here, we present a brief version of the model with money; the in-depth version with formal statements and proofs is provided in the web appendix.

Because money and the yield of recently matured assets are equally good as media of exchange, money and short term assets will be perfect substitutes in the CM and in the OTC. We can combine their quantities into a measure of real balances, so that the description of trade in section 3.2 is unchanged; we only have to define $z \equiv \varphi m + a_1$, where m denotes a buyer's

[21] It is important to highlight that Vayanos and Weill (2008) define liquidity in a slightly different way than we do. In that paper, liquidity is defined as the ease with which agents can find buyers for their assets. Here, assets are liquid primarily because they can help agents facilitate trade in the anonymous LW market (money and short term assets). However, long term assets also have indirect liquidity properties, since they can help agents acquire liquid assets in the OTC market. In fact, the latter notion of liquidity (that of long term assets) is quite close to the one employed by Vayanos and Weill: the liquidity of long term assets is directly determined by the ease with which they can be traded (for money and maturing assets) in the OTC market.

individual money holdings, and analogously for $\tilde{z}$. The description of optimal choices in 3.3 is almost identical, too, with the caveat that people will only demand positive amounts of both money and short term assets if $\varphi/\hat{\varphi} = \psi_1$. If that is the case, $\hat{z} \equiv \hat{\varphi}\hat{m} + \hat{a}_1$.

The equilibrium level of real balances is given by $Z \equiv \varphi M + A_1$ satisfying the money demand equation in either Region 1 or 5. As long as $\varphi > 0$ we have a monetary equilibrium. In the steady state of a monetary equilibrium, expected inflation $\varphi/\hat{\varphi}$ must equal the growth rate of the money supply $1 + \mu$, and consequently we must have $\psi_1 = 1 + \mu$ for the price of short term assets. The price of long term assets is now given by $\psi_2 = \beta(1 + \rho)\psi_1$ as before, with the same formula for ρ as in (26), but with q_2 now defined in terms of (Z, A_2) instead of (A_1, A_2).

If inflation is too high, however, money will not be valued as short term assets would be superior both as a store of value and as a medium of exchange. This situation will obtain if $Z = A_1$, in which case we would be back in the analysis of Section 3. We can formally express this upper bound on inflation as $\bar{\mu}(A_1, A_2) = \psi_1^{NM} - 1$, where ψ_1^{NM} denotes the price of short term assets in the nonmonetary equilibrium of Section 3. Clearly, if $A_1 \geq q^*$ so that $\psi_1^{NM} = \beta$, no monetary equilibrium can exist.

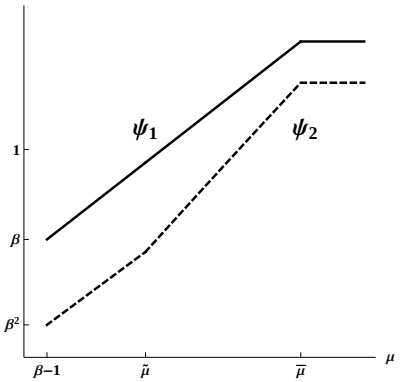

Figure 5: Equilibrium asset prices as functions of inflation. $\tilde{\mu}$ denotes the threshold beyond which long term assets are scarce in OTC trade, and $\bar{\mu}$ is the upper bound on inflation beyond which money is not valued.

The effect of monetary policy on equilibrium asset prices is easy to describe. Money and short term assets are perfect substitutes, so the price of short term assets increases in inflation as long as equilibrium remains monetary. The price of long term assets also increases in inflation; first, indirectly through the price of short term assets, and second, directly in Region 5 equilibria because inflation increases ρ (through decreasing q_2). Long term assets are imperfect substitutes to money. These results are depicted in Figure 5.

Finally, using the money demand equation, we can prove that $1 \leq 1 + \rho < (1 + \mu)/\beta$. This fact establishes that the real term premium $\psi_2^{-0.5} - \psi_1^{-1}$ is positive in the steady state of any monetary equilibrium. In order to study the effect of inflation on the term premium, focus on the regions of monetary equilibrium, and, for simplicity, consider the case of plentiful A_2 (the argument for Region 5 equilibria is slightly more complicated since it involves the derivative

of ρ with respect to μ). In this region, we know that $\tau_{1,2} = [\beta(1+\mu)]^{-1/2} - (1+\mu)^{-1}$. Clearly, as $\mu \to \beta - 1$, we have $Z \to q^*$, and, consistent with the discussion above, $\tau_{1,2} \to 0$. Moreover, one can easily verify that $\partial\tau_{1,2}/\partial\mu$ is positive *iff* $\mu < 4\beta - 1$. For reasonable (not too small) values of β, we have $\bar{\mu} < 4\beta - 1$ anyway (where $\bar{\mu}$ is the upper bound on inflation to support monetary equilibria, defined earlier in this section); thus, in monetary equilibria, $\tau_{1,2}$ is increasing in μ. This result is quite intuitive. Inflation increases the prices of (and reduces the interest rates on) assets of both maturities, which means that the sign of $\partial\tau_{1,2}/\partial\mu$ might be ambiguous. However, the effect of inflation on ψ_1 (or r_1) is stronger because short term assets are closer substitutes to money, implying that $\partial\tau_{1,2}/\partial\mu > 0$.

4.2 Equilibrium in the Model with N Maturities

Extending the baseline model to include longer-term assets is straightforward. Here, we provide a brief, verbal description of the results and relegate the detailed analysis to the accompanying web appendix.

One issue that arises with $N > 2$, is that there are many combinations of long term asset portfolios that a C-type can sell in order to obtain additional liquidity in the OTC market. We choose to not place any restrictions on which assets can be traded for liquidity. That is, we assume that in any OTC meeting the C-type can exchange any portfolio of long term assets for a portfolio of liquid assets (i.e. money and the yield of assets that mature in the current period). Therefore, even though $N > 2$, the interesting distinction is still between assets that mature now (and are therefore liquid) and assets that mature in future periods, and the results (summarized in Proposition 4 of the web appendix) are qualitatively similar to the ones in the $N = 2$ case. In particular, one-period assets are "in a class of their own", since they are the only assets that are (direct) substitutes to money. Hence, in any monetary equilibrium we must have $\psi_1 = 1 + \mu$. The price of longer term assets, ψ_i, $i \geq 2$, always carries a liquidity premium because these assets will eventually also become short term assets in future periods. Moreover, if the supply of longer term assets is relatively scarce, ψ_i will also contain an *indirect* liquidity premium which reflects the assets' property to be traded for liquid assets in the OTC market, and this premium is increasing in inflation and decreasing in the supply of long term assets.

In the web appendix, we show that the model with $N > 2$ maturities delivers an upward sloping yield curve for all $i = 1, ..., N$, regardless of the region of equilibrium. This result emerges even though any two assets with lifetime $i, j \geq 2$ are qualitatively similar, in that neither of them can serve as a direct substitute to money. Nevertheless, assets with maturity $i \geq 2$ are, in a sense, still more liquid than assets with maturity $i + 1$ because the former will become one-period assets (and perfect substitutes to money) *earlier* than the latter.[22]

[22] It would be possible to extend to the model in a way that makes longer-term assets imperfect substitutes. If regular access to the CM was not guaranteed, or if new assets were not issued in every CM, an asset buyer (in the

4.3 Quantitative Implications

While the contribution of our paper is theoretical, we want to check whether our proposed mechanisms are consistent with empirical magnitudes. It is not obvious how to best approach this: the assumption that one-period assets always mature just in time to take advantage of consumption opportunities makes our analysis transparent but is too strict for a calibration. Furthermore, our OTC market is very stylized: agents trade at most once, and with a single counterparty, and there are no dealers. We believe that this is the best way to capture the *essential* frictions associated with reallocating asset portfolios (which dealers can only mitigate, not eliminate), but it naturally constrains any empirical analysis.

Despite these limitations, our model is broadly consistent with the data. With some algebra, assuming the liquidity premia θ and ρ are small, we can approximate the term premium between assets of any two maturities by:

$$\tau_{n,m} \approx \frac{m-n}{mn} \cdot \theta \quad \text{for } m > n.$$

There are two possible approaches to estimating the direct liquidity premium θ. If we take the monetary version of our model literally, where short term assets are perfect substitutes to money, then $\theta = i - \rho$, where $i \equiv (1+\mu-\beta)/\beta$, and $i-\rho$ would be the limiting interest rate on a very long-term nominal asset (as follows from Sections 2 and 3 of the web appendix). Say that $i - \rho = 5\%$; in this case, the yield difference between one- and two-year discount bonds is 250 basis points if the calibration period is one year, and 62 basis points if the calibration period is one quarter.

However, in reality, short term assets are not *perfect* substitutes to money. In order to avoid taking a stand on exactly how close their substitutability is, we can use the non-monetary version of the model, and derive an estimate for θ using the on-the-run premium implied by our analysis in Section 3.6.3, as follows. In an environment where liquidity is not too scarce, $u'(q_2)$ is close to 1 and therefore we can approximate $\theta \approx \ell(1-\lambda)\left[u'(q_1)-1\right]$ (where we also assume $f = \ell$). Similarly, if $u'(q_2)$ is close to $u'(q_1)$, then the on-the-run premium $\omega \equiv \psi_1/\psi_o - 1$ can be approximated by $\omega \approx (1-\lambda)\left[u'(q_1)-1\right]$. The approximation errors go in opposite directions, reducing bias but adding uncertainty. Substituting, we can link the term premium and the on-the-run premium:

$$\tau_{n,m} \approx \frac{m-n}{mn} \cdot \ell \cdot \omega$$

CM or as an N-type in the OTC) might prefer 2-period assets over even longer-term ones, not only because she expects that the former will mature closer to potential liquidity needs, but also because if she needs to sell them in the next period in the OTC, she expects that the potential buyer of these assets will in turn value the shorter-term assets more. Our model does not account for this type of effect because of the discrete time structure and the existence of the CM, which in effect washes out all previous trading histories.

Vayanos and Weill (2008) report estimates of the on-the-run premium of 30-60 basis points. We set ℓ, the frequency of liquidity shocks, to 0.5. If the length of a period is one year, then the yield difference between one- and two-year assets is between 7.5 and 15 basis points, and if the length of a period is one quarter, then it falls to between 2 and 4 basis points.

The numbers from the two approaches are so divergent because the empirical limitations of our model pull in opposite directions. The assumption that OTC trade can only happen once and only with one counterparty implies a large on-the-run premium – unless liquidity is plentiful or the C-type has a high bargaining power. Both of these would imply a small term premium but also overall interest rates that are unrealistically low. If we instead use the information contained in the overall level of interest rates, the assumption that short term assets always mature just in time implies a large term premium and a large on-the-run premium. Certainly, our paper is a theory paper chiefly concerned with getting directions right, not magnitudes; but between the two, we are more comfortable matching the term premium, as our OTC market is really very stylized. At any rate, our numbers do bracket Cochrane's (1999) estimate of 32 basis points for the term premium between one- and two-year discount bonds.

5 Conclusions

Liquidity preference is often proposed as a resolution to the well-documented empirical failures of the expectations hypothesis of the term structure. This paper provides a theoretical basis for this preference. We extend the standard monetary-search model of Lagos and Wright (2005) to include assets of different maturities. Short term assets mature in time to take advantage of random consumption opportunities. Long term assets cannot be used directly to purchase consumption, but agents may liquidate them in a secondary asset market. To make things interesting and realistic we follow the influential work of Duffie, Gârleanu, and Pedersen (2005) and model this market as an OTC market characterized by search and bargaining frictions.

Our model provides a general framework within which one can think of asset liquidity, and how the latter can affect asset prices. The model delivers three results which are consistent with empirical findings. First, in equilibrium, long term assets have higher rates of return to compensate agents for the cost associated with liquidating these assets in the frictional OTC market. Second, our model predicts that the yield curve will be steeper for assets that trade in less liquid secondary markets. Finally, our model provides a simple and intuitive explanation for the "on-the-run phenomenon". In particular, we find that freshly issued assets will sell at higher prices than previously issued assets that mature on nearby dates, because sellers of "off-the-run" assets, who are in need for liquidity, may have a low bargaining power in the OTC market and are therefore willing to sell assets at lower prices.

A Appendix: Proofs of Main Lemmas and Propositions

Proof. Proof of Lemma 3.

Consider the derivatives of the objective function with respect to $\hat{z}$ and $\hat{a}_2$, i.e. equations (17)-(23). To obtain these conditions we substitute the appropriate solution to the bargaining problem (depending on the region in question) into (16), and we differentiate with respect to $\hat{z}$ or $\hat{a}_2$.

As an illustration, consider Region 2. Recall that in this region, $\hat{z} < q^* - \tilde{z}$, $\hat{a}_2 > \bar{a}(\hat{z}, \tilde{z})$, but $\tilde{a}_2 < \bar{a}(\tilde{z}, \hat{z})$. Based on this information, we have $\chi = \bar{a}(\hat{z}, \tilde{z})$, $\zeta = \tilde{z}$, $\tilde{\chi} = \tilde{a}_2$, and $\tilde{\zeta} = \zeta^a(\tilde{z}, \tilde{a}_2)$. Substituting these terms into the objective function implies that

$$\begin{aligned}\beta^{-1} J^2(\hat{z}, \hat{a}_2) = &-\frac{\psi_1 \hat{z} + \psi_2 \hat{a}_2}{\beta} + f \left\{u\left(\hat{z} + \tilde{z}\right) - \beta\psi_1 \bar{a}(\hat{z}, \tilde{z})\right\} \\ &+ (\ell - f)\, u(\hat{z}) + f \left\{[\hat{z} - \zeta^a(\tilde{z}, \tilde{a}_2)] + \beta\psi_1 \tilde{a}_2\right\} + (1 - \ell - f)\, \hat{z}.\end{aligned}$$

The remaining derivations follow exactly the same steps.

Notice that we can solve $J_1^i = 0$, $i = 1, ..., 5$, with respect to the term $\hat{\psi}_1$. This will yield the demand for real balances as a function of their holding cost. For future reference, it is important to highlight that the demand for real balances is in fact continuous on the boundaries 1-2, and 1-5.[23] Similarly, we can solve $J_2^i = 0$, $i = 1, ..., 5$, with respect to $\psi_2/(\beta\hat{\psi}_1)$, in order to obtain the demand for long term assets. It can be easily verified that this function is continuous on the boundaries 1-2, 2-5, 2-3, and 4-5.

Some preliminary facts about the objective function $J : \mathbb{R}^3_+ \to \mathbb{R}$:

Fact 1: J is continuous everywhere.

Proof: The solution to the OTC bargaining problem is continuous. One of the three constraints $\zeta \leq \tilde{z}$, $\zeta \leq q^* - z$, and $\chi \leq a_2$ must bind, together with equation (10). Each of these is linear in the choice variables. Therefore, J is continuous.

Fact 2: J is differentiable within each of the five regions defined above.

Proof: As above, one of the constraints must bind together with equation (10). Each of these is differentiable in the choice variables, and within a region of J, the binding constraint does not switch. Furthermore, J is differentiable on those boundaries where both FOCs are continuous (see above).

Fact 3: J is strictly concave in the first argument (real balances) whenever $z < q^*$.

Proof: As J is continuous everywhere and differentiable within each region, J_1 is defined everywhere except at a finite number of boundary crossings. We need to show that J_1 is decreasing as a function of $\hat{z}$ within each region, and that $J_{1-} \geq J_{1+}$ on each boundary, where "–" denotes the left derivative and "+" denotes the right derivative.

[23] The demand for real balances is also continuous on the boundaries of the Regions 1-3 and 4-5 if $\tilde{a}_2 \geq \bar{a}(\tilde{z}, q^* - \tilde{z})$, in which case Region 2 does not exist.

That J_1 is strictly decreasing in $\hat{z}$ within Regions 1-3 follows immediately from equations (17)-(19), and the fact that u' is strictly decreasing. In Regions 4 and 5, showing that J_1 is decreasing in $\hat{z}$ is less obvious. In Region 5 (where $\hat{z}+\zeta < q^*$), we have

$$J_1^5 = -\psi_1 + \beta\ell\left[u'(\hat{z})-1\right] + \beta\lambda f\frac{u'(\hat{z}+\zeta)-u'(z)}{(1-\lambda)u'(\hat{z}+\zeta)+\lambda}.$$

Since ζ satisfies equation (14), applying total differentiation to that equation yields

$$\frac{d\zeta}{d\hat{z}} = (1-\lambda)\frac{u'(\hat{z})-u'(\hat{z}+\zeta)}{(1-\lambda)u'(\hat{z}+\zeta)+\lambda}.$$

Consequently,

$$\begin{aligned}\frac{\partial J_1^5}{\partial z} = \frac{\beta}{\left[(1-\lambda)u'(\hat{z}+\zeta)+\lambda\right]^3}\Big\{ & f\lambda\left[(1-\lambda)u'(\hat{z})+\lambda\right]^2 u''(\hat{z}+\zeta) \\ & + \left[(\ell-f)\lambda+\ell(1-\lambda)u'(\hat{z}+\zeta)\right]\left[(1-\lambda)u'(\hat{z}+\zeta)+\lambda\right]^2 u''(\hat{z})\Big\}.\end{aligned}$$

Since $u''(\,\cdot\,)<0$, the entire term $\partial J_1^5/\partial\hat{z}<0$. In Region 4, the only addition is a term involving $u'(\,\cdot\,)$, which is clearly decreasing too. Hence, J_1^4 is decreasing in $\hat{z}$ as well.

As we discussed above, J_1 is continuous across all the boundaries of the various regions, except the boundaries 2-3, 3-4, 4-5, 2-5, and the crossing 2-4. With some algebra, one can check that $J_1^2 < J_1^3$, $J_1^3 < J_1^4$, $J_1^4 < J_1^5$, and $J_1^2 < J_1^5$, across the respective boundaries. Also, $J_1^3 > J_1^5$ at the crossing 2-3-4-5, establishing the chain $J_1^2 < J_1^5 < J_1^3 < J_1^4$ at this crossing. Consequently, J is concave in $\hat{z}$ throughout.

Fact 4: J is concave in the second argument (long term assets), strictly in Regions 4 and 5.

Proof: As J is continuous everywhere and differentiable within each region, J_2 is defined everywhere except at a finite number of boundary crossings. We need to show that J_2 is decreasing as a function of $\hat{a}_2$ within each region (strictly, in Regions 4 and 5), and that $J_{2-} \geq J_{2+}$ on each boundary, where "–" denotes the left derivative and "+" denotes the right derivative. In Regions 1-3, J_2^i is constant, hence weakly concave. We now show that J_2^i is strictly decreasing in $\hat{a}_2$ within Regions 4 and 5. Applying total differentiation to equation (14), yields

$$\frac{\partial\zeta}{\partial a_2} = \frac{\beta\hat{\psi}_1}{(1-\lambda)u'\left(\hat{z}+\zeta\right)+\lambda}.$$

Since this expression is clearly positive, and u' is strictly decreasing, it follows that $\partial J_2^i/\partial\hat{a}_2<0$, for $i=4,5$.

Next, using the definitions of the regions, one can see that J_2 is continuous across the boundary 1-5, but not the boundaries 2-5 or 3-4. The term $u'(\hat{z}+\zeta)[(1-\lambda)u'(\hat{z}+\zeta)+\lambda]^{-1}$ is greater than 1 in Regions 4 and 5, because $\hat{z}+\zeta<\min\{\hat{z}+\tilde{z},q^*\}$ (by definition of Regions 4 and 5), and

therefore $u'(\,\cdot\,) > 1$.

Fact 5: J is weakly concave everywhere.

Proof: We need to show that J_2 is non-increasing as a function of $\hat{z}$ within each region, and across boundaries. First, J_2 depends on $\hat{z}$ only in Regions 4 and 5. There, ζ is strictly increasing in $\hat{z}$, therefore $u'(\hat{z}+\zeta)$ is strictly decreasing, and so is $u'(\hat{z}+\zeta)[(1-\lambda)u'(\hat{z}+\zeta)+\lambda]^{-1}$.

Now, the only boundaries where J_2 is not a continuous function of $\hat{z}$ are the boundaries of Regions 3 and 4, and 2 and 5, which are downward sloping in $(\hat{z}, \hat{a}_2)$-space. On these boundaries, $J_{2-} > J_{2+}$ (see Fact 4). This is sufficient because an infinitesimal increase in $\hat{z}$ has the same effect as an infinitesimal increase in $\hat{a}_2$ (the definition of J_{2+}), and vice versa, as the boundaries are downward sloping in $(\hat{z}, \hat{a}_2)$-space.

We conclude that J_2 is weakly decreasing as a function of $\hat{z}$, therefore J is submodular (real balances and long term assets are strategic substitutes). As J is also weakly concave in each argument, it is weakly concave overall.

Proof of the statement of the Lemma:

a) The fact that $\nabla J = \mathbf{0}$ at the solution follows from the fact that J is weakly concave overall and differentiable within each region. So if the optimal choice $(\hat{z}, \hat{a}_2)$ is within a region, the first-order conditions must hold.

b) In Regions 1-3, demand for real balances is strictly decreasing, so $\hat{z}$ is unique as long as $\psi_1 > \beta$. But any $\hat{a}_2$ in Regions 1-3 satisfies $J_2^i = 0$, $i = 1, 2, 3$, and the fact that $\psi_2 = \beta\hat{\psi}_1$ rules out Regions 4 and 5. To see this point, notice from (23) that for any $(\hat{z}, \hat{a}_2)$ in the interior of these regions, $\psi_2 = \beta\hat{\psi}_1$ implies $\beta J_2^i > 0$, for $i = 4, 5$.

c) The fact that $\psi_2 > \beta\hat{\psi}_1$ rules out the interior of Regions 1-3 or the boundary 1-5. To see why, notice from (22), that for any $(\hat{z}, \hat{a}_2)$ in the regions in question, $\psi_2 > \beta\hat{\psi}_1$ implies $J_2^i < 0$, for $i = 1, 2, 3$. □

Proof. Proof of Lemma 4.

The equilibrium objects q_1, q_2, χ, and ζ are all deterministic functions of $Z = A_1$, so it suffices to focus on ψ_1, and ψ_2. If $A_1 = q^*$, then $q_1 = a_2 = q^*$ as well as $\chi = \zeta = 0$ and $\psi_1 = \beta$ and $\psi_2 = \beta^2$. If $A_1 < q^*$, we have $\psi_1 > \beta$. Consequently, parts (b) and (c) of Lemma 3 apply, and an optimal $(\hat{z}, \hat{a}_2)$ exists and $\hat{z}$ is unique. The object ψ_1 must be such that $\hat{z} = Z = A_1$ satisfies the demand for real balances, $J_1 = 0$.

Finally, set $\hat{a}_2 = A_2$. The assumption $A_1 \geq A_2 + \chi(A_1, A_1, A_2)$ guarantees that agents never need to sell assets in the CM; N-types held two-period assets A_2 at the end of the preceding period, which become one-period assets in the given period, and obtain χ more in the OTC market if they are matched. C-types and unmatched N-types will enter the CM with less than $A_2 + \chi$ one-period assets, so every agent can obtain the symmetric quantity of short term assets, A_1, by buying newly issued ones and not by selling previously-issued ones.

Additionally, if the parameters of the model satisfy inequality (24), then the equilibrium

must be in Regions 1 or 5, as described in the text. Now examine the demand function for long term assets (equations (22) and (23)). It is constant in Regions 1 and strictly decreasing in $\hat{a}_2$ in Region 5 (also see the proof of Lemma 3, Fact 4), and is continuous on the boundary of Regions 1 and 5. If (A_1, A_2) lies in the interior of Region 5, then $\psi_2 > \beta\psi_1$ is unique. If (Z, A_2) lies in the interior of Region 1 or on the boundary of Regions 1 and 5, then $\psi_2 = \beta\psi_1$, which is unique. □

Proof. Proof of Proposition 1.

If $A_1 \geq q^*$, then $q_1 = a_2 = q^*$ is an equilibrium with $\psi_1 = \beta$ and $\psi_2 = \beta^2$. OTC bargaining yields $\chi = \zeta = 0$. □

Proof. Proof of Propositions 2 and 3.

Recall that $A_1 < q^*$ is a maintained assumption throughout.

Case 1: Let $A_2 \geq \bar{A}_2(A_1)$. Then the equilibrium is in Region 1 in (A_2, Z)-space. By equation (22), the only solution to $J_2 = 0$ in Region 1 is $\psi_2 = \beta\psi_1$. Furthermore, Region 1 is defined by the branch of the OTC bargaining solution where $\zeta = q^* - z$, so on aggregate, $q_2 = Z + \zeta(Z, Z, A_2) = q^*$. Therefore the only solution to $J_1 = 0$ is $\psi_1 = \beta(1 + \theta)$ with θ defined in 25.

Case 2: Let $A_2 < \bar{A}_2(A_1)$. Then the equilibrium is in Region 5. In this region the first-order conditions $J_1^5 = 0$ (demand for RB) and $J_2^5 = 0$ (demand for long term assets) apply, evaluated at aggregate quantities, which are exactly the equations in the statement together with (25) and (26). The description of $\tilde{q}$ follows from substituting these asset pricing equations into the definition of $\tilde{q}$ from Definition 1. This definition in turn is derived from the OTC bargaining solution on the branch where long term assets are scarce, given in equation (14). □

B Appendix: Equilibrium Production in the LW market

Consider the equilibrium quantities traded in the LW market, assuming that $A_1 < q^*$.

Proposition 3. *The equilibrium value of q_1 is always equal to A_1, and the equilibrium value of q_2:*

Case 1: *If $A_2 \geq \bar{A}_2(A_1)$, which is only admissible if also $A_1 > \bar{A}_1$, then $q_2 = q^*$.*

Case 2: *If $A_2 < \bar{A}_2(A_1)$, then $q_2 = \tilde{q}(A_1, A_2)$ which is described by the solution to:*

$$(1-\lambda)\left[u(\tilde{q}) - u(A_1)\right] + \lambda\left(\tilde{q} - A_1\right) = \ldots$$
$$\ldots \beta A_2 \left(1 + \left[\ell - \frac{\lambda f}{(1-\lambda)u'(\tilde{q}) + \lambda}\right]\left[u'(A_1) - 1\right] + \frac{\lambda f[u'(\tilde{q}) - 1]}{(1-\lambda)u'(\tilde{q}) + \lambda}\right).$$

Proof. Jointly proven with Proposition 2, in Appendix A. □

The results demonstrated in Proposition 3 are also very intuitive. Agents who did not match in the OTC have to rely exclusively on their own real balances. Hence, q_1 will always coincide with $Z = A_1$. The equilibrium quantity q_2 represents the amount of goods that the buyer can afford to purchase in the LW market, when she has previously traded in the OTC market. Hence, whenever equilibrium lies in Region 1, we have $q_2 = q^*$. In contrast, if equilibrium lies in the regions of scarcity of A_2 in OTC trade (the admissible section of Region 5), the buyer will not be able to afford the first-best, and $q_2 < q^*$.

Regarding the comparative statics of $\tilde{q}$, it turns out that this variable increases less than one-to-one with the supply of real balances. The reason for this is the bargaining process in the OTC market. Say a buyer considers whether carry one additional unit of real balances A_1 into the OTC market. She knows that she could purchase the same amount of the N-type's real balances as before and ultimately end up with one more unit of goods q_2. However, this OTC purchase is costly if she does not have all the bargaining power, as she is giving up long term assets at a relatively low price. Consequently, due to concave utility, she will split the benefit between purchasing more LW goods ($d\tilde{q}/dA_1 > 0$) and buying fewer real balances at a premium ($d(\tilde{q} - A_1)/dA_1 < 0$).

Similarly, if a buyer carries one additional unit of long term assets A_2 into the OTC market, she will not spend all of it on real balances, because these additional real balances ($d\tilde{q}/dA_2 > 0$) move her down her demand curve in the LW market, which reduces her valuation of marginal liquidity and makes her retain some of her long term assets rather than selling them at a discount relative to their continuation value ($d\tilde{q}/dA_2 < \psi_1$).

References

ANDOLFATTO, D., A. BERENTSEN, AND C. WALLER (2014): "Optimal disclosure policy and undue diligence," *Journal of Economic Theory*, 149, 128–152.

ANDOLFATTO, D., AND F. M. MARTIN (2013): "Information disclosure and exchange media," *Review of Economic Dynamics*, 16(3), 527–539.

BACKUS, D. K., A. W. GREGORY, AND S. E. ZIN (1989): "Risk premiums in the term structure: Evidence from artificial economies," *Journal of Monetary Economics*, 24(3), 371–399.

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.

BOEL, P., AND G. CAMERA (2006): "Efficient monetary allocations and the illiquidity of bonds," *Journal of Monetary Economics*, 53(7), 1693–1715.

CHALLE, E., F. LE GRAND, AND X. RAGOT (2013): "Incomplete markets, liquidation risk, and the term structure of interest rates," *Journal of Economic Theory*, 148(6), 2483–2519.

COCHRANE, J. H. (1999): "New facts in finance," Discussion paper, National Bureau of Economic Research.

DUFFIE, D., N. GÂRLEANU, AND L. H. PEDERSEN (2005): "Over-the-Counter Markets," *Econometrica*, 73(6), 1815–1847.

ENGLE, R. F., M. J. FLEMING, E. GHYSELS, AND G. NGUYEN (2012): "Liquidity, volatility, and flights to safety in the US Treasury market: Evidence from a new class of dynamic order book models," *FRB of New York Staff Report*, (590), 2013–20.

GARBADE, K. (2007): "The Emergence of "Regular and Predictable" as a Treasury Debt Management Strategy," *Economic Policy Review*, 13(1).

GARBADE, K., AND J. INGBER (2005): "The Treasury auction process: Objectives, structure, and recent adaptations," *Current issues in economics and finance*, 11(2).

GEROMICHALOS, A., AND L. HERRENBRUECK (2012): "Monetary Policy, Asset Prices, and Liquidity in Over-the-Counter Markets," Working paper, University of California, Davis.

——— (2015): "The Strategic Determination of the Supply of Liquid Assets," *mimeo*.

GEROMICHALOS, A., J. M. LICARI, AND J. SUAREZ-LLEDO (2007): "Monetary Policy and Asset Prices," *Review of Economic Dynamics*, 10(4), 761–779.

GEROMICHALOS, A., AND I. SIMONOVSKA (2014): "Asset liquidity and international portfolio choice," *Journal of Economic Theory*, 151, 342–380.

GREEN, R. C. (1993): "A simple model of the taxable and tax-exempt yield curves," *Review of Financial Studies*, 6(2), 233–264.

GÜRKAYNAK, R. S., B. SACK, AND J. H. WRIGHT (2010): "The tips yield curve and inflation compensation," *American Economic Journal: Macroeconomics*, pp. 70–92.

GÜRKAYNAK, R. S., AND J. H. WRIGHT (2012): "Macroeconomics and the term structure," *Journal of Economic Literature*, 50(2), 331–367.

HEATON, J., AND D. LUCAS (1992): "The effects of incomplete insurance markets and trading costs in a consumption-based asset pricing model," *Journal of Economic Dynamics and Control*, 16(3), 601–620.

HU, T.-W., AND G. ROCHETEAU (2013): "On the coexistence of money and higher-return assets and its social role," *Journal of Economic Theory*, 148(6), 2520–2560.

JACQUET, N. L., AND S. TAN (2012): "Money and asset prices with uninsurable risks," *Journal of Monetary Economics*, 59(8), 784–797.

KALAI, E. (1977): "Proportional Solutions to Bargaining Situations: Interpersonal Utility Comparisons," *Econometrica*, 45(7), 1623–30.

KRISHNAMURTHY, A., AND A. VISSING-JORGENSEN (2012): "The aggregate demand for treasury debt," *Journal of Political Economy*, 120(2), 233–267.

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

——— (2011): "Asset Prices, Liquidity, and Monetary Policy in an Exchange Economy," *Journal of Money, Credit and Banking*, 43, 521–552.

LAGOS, R., AND G. ROCHETEAU (2008): "Money and capital as competing media of exchange," *Journal of Economic Theory*, 142(1), 247–258.

LAGOS, R., AND R. WRIGHT (2005): "A Unified Framework for Monetary Theory and Policy Analysis," *Journal of Political Economy*, 113(3), 463–484.

LESTER, B., A. POSTLEWAITE, AND R. WRIGHT (2012): "Information, liquidity, asset prices, and monetary policy," *The Review of Economic Studies*, 79(3), 1209–1238.

LUCAS, ROBERT E, J. (1978): "Asset Prices in an Exchange Economy," *Econometrica*, 46(6), 1429–45.

MISHKIN, F. S. (2007): *The economics of money, banking, and financial markets*. Pearson education.

NOSAL, E., AND G. ROCHETEAU (2013): "Pairwise trade, asset prices, and monetary policy," *Journal of Economic Dynamics and Control*, 37(1), 1–17.

PIAZZESI, M. (2010): "Affine term structure models," *Handbook of financial econometrics*, 1, 691–766.

PIAZZESI, M., AND M. SCHNEIDER (2007): "Equilibrium yield curves," in *NBER Macroeconomics Annual 2006, Volume 21*, pp. 389–472. MIT Press.

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

ROCHETEAU, G., AND R. WRIGHT (2005): "Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium," *Econometrica*, 73(1), 175–202.

SALYER, K. D. (1990): "The term structure and time series properties of nominal interest rates: Implications from theory," *Journal of Money, Credit and Banking*, 22(4), 478–490.

SINGLETON, K. J. (2009): *Empirical dynamic asset pricing: model specification and econometric assessment*. Princeton University Press.

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.

WARGA, A. (1992): "Bond returns, liquidity, and missing data," *Journal of Financial and Quantitative Analysis*, 27(4).

WILLIAMSON, S. D. (2012): "Liquidity, monetary policy, and the financial crisis: A New Monetarist approach," *The American Economic Review*, 102(6), 2570–2605.

——— (2013): "Scarce Collateral, the Term Premium, and Quantitative Easing," .