Pricing and Liquidity in Over-The-Counter Markets*

Oliver Randall†

Emory University - Goizueta Business School

January 12th, 2017

Abstract

I show that if dealers are averse to holding inventory, then prices, liquidity, and dealers' inventory positions depend on inventory costs in negotiated over-the-counter markets. The solution to my dynamic equilibrium model rationalizes the following stylized facts in the US corporate bond market: (i) a reduction in dealers' inventories during crises; (ii) a reduction in average trade size since the onset of the financial crisis and tighter regulatory environment; (iii) better prices for customers to buy than sell in the financial crisis; (iv) a generally negative relationship between transaction costs and trade size. The results inform debate on the Volcker Rule.

EFM Classification: 340, 360, 520

*The author would like to thank Patrick Augustin, Joel Hasbrouck, Esben Hedegaard, Rustom Irani, Hanh Le, Anthony Lynch, Albert Menkveld, Emiliano Pagnotta, Lasse Pedersen, Or Shachar, Marti Subrahmanyam, and seminar participants at Emory University, New York University, the Federal Reserve Board, HEC, ESSEC, Tilburg, NHH, BI, Stockholm School of Economics, University of Virginia, University of Melbourne, University of New South Wales, the All-Georgia Finance conference, and University of Illinois at Chicago for helpful comments and suggestions. All errors are of course my own.

†Goizueta School of Business, Emory University, 1300 Clifton Road, Atlanta, GA 30322, Email: oliver.randall@emory.edu, Phone: (404) 727-0482, Fax: (404) 727-5238.

1 Introduction

Dealers in the US corporate bond market have reduced their inventories substantially since October 2007[1]. This coincided with the start of large write-downs on the balance sheets of several banks (Citi, Merrill Lynch, UBS),[2] whose dealer businesses had substantial market share in liquidity provision in the US corporate bond market. In the financial press[3] this reduction in inventory has been attributed to a reduction in dealer risk appetite during the financial crisis of 2007-09. During that period insurance companies were net buyers of corporate bonds.[4] In the subsequent sovereign debt crisis, dealers again reduced their aggregate inventories of corporate bonds. But Weill (2007) and Lagos, Rocheteau and Weill (2011) show that if dealers are risk-neutral and customers become more risk averse in a crisis, dealers should "lean against the wind" and stock up on risky assets to sell them back to customers at a profit, once customers' risk tolerance has returned to pre-crisis levels. Though those models are applicable to other crises and other markets, they were not designed for the US corporate bond market during the financial crisis, nor for the subsequent sovereign debt crisis, when dealers reduced their aggregate inventories, rather than increasing them. The data suggests that these crises affected dealers more than their customers, and Randall (2015) shows that it was not just the absolute level of dealer risk appetite, but also the level relative to their customers, which determined which set of agents built or reduced inventory in that period. In that sense, customers took the role of leaning against the wind in these crises in the corporate bond market. Since the dealers in the customer-crisis models, and in other over-the-counter (OTC) trading models of Lagos and Rocheteau (2009), and Duffie, Garleanu, Pedersen (2005, 2007), are risk neutral, the price at which they trade with customers does not depend on their inventory positions, the riskiness of the traded asset, nor the time-varying cost to the dealers of bearing inventory risk. But another empirical stylized fact for the US corporate bond market is that prices and liquidity changed dramatically during the financial crisis, with asymmetric effects for customer purchases and sales. Figure 2 shows the starkness of this price asymmetry in the financial crisis: on average customers who were prepared to buy from a dealer market which was looking to offload inventory could buy at just a small markup to the inter-dealer price, but prices offered to customers who were selling were marked down severely. Figure 3 also shows that trade size has decreased a

[1] See Figure 1, Choi and Shachar (2014), and "Digging into dealer inventories" (September 11th 2013, FT.com). Though the extent of the deleveraging is hard to estimate, all these studies agree dealer inventories are lower than their pre-crisis peak.

[2] See "Restoring Financial Stability: How to Repair a Failed System" (2009)

[3] e.g. "Slimmer bond inventories as dealers reduce risk", November 8th, 2011 (FT.com)

[4] See Randall (2015).

lot since the onset of the financial crisis. All this evidence is consistent with dealers bearing a time-varying cost of holding inventory, which increases markedly in crisis periods.

Though there are already theoretical models highlighting the effects of costly dealer inventory for anonymous trade in the equity markets, such as Stoll (1978) and Amihud and Mendelson (1980), they do not incorporate the bargaining that plays out in OTC markets where agents' identities are known when trade is negotiated. So they cannot explain the empirical fact that larger trades tend to get better prices in the corporate bond market, as we see in Figure 4. In this paper I develop a model which can capture the aggregate time series facts of the US corporate bond market described above, and produce testable implications for liquidity and dealer inventory in the cross-section at trade level, which are confirmed in Randall (2015). Though the assumptions of the model are motivated by the corporate bond market, it can be applied to any OTC market.

In my theoretical model, trading of a risky asset occurs in a hybrid market, consisting of secondary trading in both a competitive inter-dealer market and in an OTC search market where customers and dealers bargain over both price and quantity, and primary trading where dealers can trade at issuance of the asset, collecting a payoff at maturity. I measure liquidity as the dealers' "markup", which is defined as the difference between the price per unit at which they trade with a customer, and the price at which the average hypothetical inter-dealer trade would have occurred, had there been one at that time. The markup can be thought of as the retail markup to the customer versus the wholesale inter-dealer price. This measure of liquidity is related to the bid-ask spread, but the benchmark price is the inter-dealer price rather than the midquote. The inter-dealer price is less arbitrary, and allows for markup asymmetry between transactions to buy versus to sell, which Figure 2 shows can be substantial. It also allows us to think of these prices and the markup as functions of the traded quantity, which, crucially, is endogenous in my model. The dealer markup is the theoretical analogue to the empirical "unique round trip" measures of Zitzewitz (2010) and Feldhutter (2012), which quantify liquidity in the US corporate bond market. Those studies focus on "paired" customer-dealer trades: ones that are accompanied by an inter-dealer trade within a few minutes, often for the exact same quantity, as the pairing dealer immediately lays off the customer trade in the inter-dealer market. The unique round-trip measure is the difference between the two prices in a paired trade. In Zitzewitz (2010) only about one third of customer-dealer trades are paired with an inter-dealer trade, though Randall (2015) shows this rate has increased more recently. My theory models the complement of this set, the unpaired trades, which make up the majority of trading.

I derive closed-form equilibrium expressions for customer-dealer prices, dealer markups, and dealers' optimal inventory. My theory emphasizes that it is dealers' time-varying cost of holding inventory, rather than inventory per se, that drives these relationships in dealer markets.

Dealers in my model are not formally risk averse, but "effectively risk averse" as in Adrian, Etula, Shin (2010) and Etula (2013), with a time-varying cost of holding inventory making their behavior observationally equivalent to those of risk averse agents, and the model analytically tractable. It can rationalize why dealers reduced their inventories of US corporate bonds in the financial crisis of 2007-09 and subsequent sovereign debt crisis, given an increase in their effective risk aversion. This can be interpreted as an increase in risk aversion itself, or a tightening of a Value-at-Risk constraint, or any increase in capital requirements or other opportunity costs of holding inventory, such that dealers behave as if they are more averse to holding risky assets. The flexibility in interpretation allows application to other OTC markets, beyond the US corporate bond market. I show how even in the presence of inventory costs, dealers' optimal inventory positions are not generally zero, which is also what we observe empirically in Figure 1 and why a static model would not suffice. Given the dynamic nature of the dealers' problem, they consider future trades with customers and other dealers, trading off the current price at which they can trade versus expected future prices to determine whether to be long or short, and by how much. All else equal, the size of their positions do, however, shrink as the cost of holding inventory increases.

In the models of dynamic, OTC, search market trading of Duffie, Garleanu, Pedersen (2005, 2007), dealer inventory is restricted to 0 or 1 unit. However, in order to understand the relationship between trade size and price, trade size must be non-trivial. Lagos and Rocheteau (2009) weaken this restriction to allow dealers to hold positive inventory of any amount. Afonso and Lagos (2015) model inter-dealer trading in the overnight federal funds market. Their model allows dealers to have more flexible utility functions, but any aversion to holding inventory is constant over time. Empirically, the time-variation in dealers' corporate bond holdings, bid-ask spread, and trade size suggest that time-varying aversion to holding inventory is an important feature in at least that OTC market, which is why it is a key feature of my model. Unrestricted trade size allows me to better understand the equilibrium relationship between price and trade size in an OTC market. So in my model dealers' choices of inventory position are unrestricted, allowing them to be long or short any quantity. I employ a friction of "trading blackouts" – that dealers can only trade at known, discrete points in time – as in Longstaff (2001, 2009) and Garleanu (2008). This emphasizes the non-trivial periods of time at which dealers bear

inventory risk for the majority of customer trades which are not immediately unwound in the inter-dealer market. To the best of my knowledge, my model is the first dynamic trading model of an OTC search market with both dealers and customers, where the dealers are averse to holding inventory and trade size is non-trivial.

The model also gives theoretical intuition for the empirical fact that larger trades in the US corporate bond market are generally associated with smaller markups, i.e. customer prices closer the inter-dealer price. This is the reverse of the relationship between trade size and liquidity for US equities which trade in limit order markets. In a limit order market, market orders which exceed the depth of the best quote walk up or down the limit order book, and so larger trades, unless split up over time, are mechanically associated with worse volume-weighted average prices. Those equity limit order markets are anonymous, but OTC markets are not. An explanation for the disparity in the liquidity-size relationship between the two markets is this difference in transparency around counterparties. Larger trades are associated with larger customers, who have higher bargaining power because dealers value the large volume of their anticipated future business. For small and medium-sized trades this bargaining power effect dominates the dealers' inventory cost effect, which is why for these size trades there is a generally negative relationship between markup and trade size. In my model too, all else equal, higher bargaining power leads to smaller markups. But it also produces the novel prediction that, controlling for customer bargaining power, larger trades are associated with larger markups, particularly when inventory costs are high. This is supported by Figure 4, which shows that for the largest trades, where bargaining power is likely to be roughly similar amongst institutional customers, the relationship between liquidity and trade size is reversed, and larger trades are associated with higher markups.

The goal of this paper is to produce an equilibrium model which can match the aggregate time series facts in the US corporate bond market. But my model gives a precise structural equilibrium relationship between dealers' inventory costs, markups, trade size, and customer bargaining power, which can be tested in the cross-section at trade level. In Randall (2015) the following predictions of these relationships are shown to hold true in the US corporate bond market: (1) Dealers reduced their inventory of high yield bonds more when their inventory costs increased relative to their customers', but reduced their inventory of investment grade bonds less, suggesting a flight to quality; (2) Markups increased when dealers' inventory costs went up; (3) This effect was generally stronger for high yield bonds than investment grades bonds, consistent with the theory that both aggregate and bond-specific risk drives the dealers' implicit

and explicit costs of holding inventory; (4) This effect was also stronger for customers with lower bargaining power, with a strikingly monotonic relationship across bargaining power groups in bond-level regression results; (5) Conditioning on customer bargaining power, this effect was more pronounced for larger trades.

I show that dealers' time-varying costs of holding inventory, and bargaining between dealers and customers over price and quantity, are both necessary to capture the aggregate stylized time-series facts of the US corporate bond market in the last several years. The model also allows me to make new, nuanced transaction-level predictions in the cross-section, for all OTC markets. These predictions are shown to hold in the US corporate bond market.

Section 2 describes the model. Section 3 describes testable hypotheses from the model, and empirical evidence. Section 4 concludes. Proofs are relegated to the Appendix.

2 The model

In this section I solve a dynamic equilibrium model of trading in a hybrid market, where there is secondary trading in a competitive inter-dealer market, between customers and dealers in an OTC search market, and, at issuance, between dealers and the issuer of the asset. I derive expressions for the prices and quantities at which customers and dealers trade, dealers' optimal inventory, and a measure of liquidity: the dealers' "markup" which is defined as the difference between the price per unit at which they trade with customers, and the price at which dealers would trade in a hypothetical inter-dealer market had it occurred.

2.1 The economy

Time is discrete, and runs forever. There is a risk-free asset paying the dealers interest at a gross rate R each period. At any point in time there is a single risky asset, with a known finite maturity. Upon maturity, the holders of the bond receive a payoff, and a new risky asset is issued. There are a continuum of customers and dealers. Customers are risk neutral, trade only once at birth, and consume all their wealth one period later. Dealers are averse to holding inventory of the risky asset, and infinitely-lived. Each period they receive interest on their cash holdings, pay an inventory cost on their risky asset position, and may or may not get to trade with at most one[5] agent: a customer, another dealer, or the issuer.

[5]Though not explicitly modeled, a paired trade, where the dealer trades in both customer-dealer and inter-dealer market simultaneaously, can be thought of as an unexpected endowment of cash to the pairing dealer, but no change in their inventory, while the other dealer gets an unexpected change in their cash and inventory positions. The dealers' cash reserves do not affect prices or quantities traded. Though the pairing dealer's inventory is unaffected,

I denote the current state of the economy at time t as s_t, whose value changes just after each round of trading according to a Markov process. I denote the history of states since the last issuance of the risky asset as $\underline{s}_t$. The time-line is displayed in Figure 5.

2.2 The dealers

There is a continuum of dealers of mass 1 with holdings of (cash, risky asset units) = $(\$_t^d, a_t^d)$ when entering time t. $\$_t^d$ includes interest accrued from the period $(t-1, t]$, which accumulates at gross rate R per period, and is paid at time t. Each dealer is risk-neutral over her cash-flows, which include trading revenues, interest on cash, inventory costs, and payoffs at the assets' maturity. But the inventory costs make her "effectively risk averse" as in e.g. Etula (2013). That is, though formally she is risk neutral in her cash flows, she behaves as if she were risk averse. This inventory cost device makes the solution analytically tractable, while still capturing the dealers' time-varying aversion to holding inventory.

At time $t = ..., -1, 0, 1, ...$ each dealer receives interest on her cash position, and pays inventory costs $f(s_t).(a_t^d)^2$ on her risky asset position, where the function $f(.)$ is taken as exogenous and strictly positive. The holding cost is quadratic in dealer inventory as in Mahanti, Nashikkar, Subrahmanyam (2008). Its convexity allows bounded solutions for trade size and inventory holdings, and the quadratic form yields a linear relationship between dealers' markups and their inventory holdings. If $f(.)$ were proportional to the variance per unit of the next price of the risky asset, then the inventory cost could be interpreted as a Value-at-Risk constraint. But since $f(.)$ is exogenous, and price variance is endogenous, this is a loose interpretation. Nonetheless, the inventory cost captures the same effect. Since customers are risk neutral, this effective risk aversion can be thought of as relating to the dealers in an absolute sense, or relative to customers: as $f(.)$ increases, the aversion of dealers to hold inventory increases, while the aversion of customers, who do not pay this cost, is unchanged. Since the dealers' role is to manage the asynchronous arrival of customer orders, they cannot diversify their portfolios to the same extent that a customer can. Also the customers are the ones who ultimately want to hold the inventory for an extended period. In the US corporate bond market, bonds are often held by long-term investors, such as insurance companies and pension funds, looking to match the duration of their long-term liabilities with long-horizon assets, collecting regular coupons to pay out to their stakeholders. This motivates the relative inventory cost in the model, with the dealers less willing to hold inventory than their customers.

aggregate dealer inventory changes by the size of the trade.

Though there is a continuum of dealers, there are a number of distinct dealer types, denoted N_D, of equal measure, identified by their inventories. At time $t..., -1, 0, 1, ...$ there are rounds of trading. Either a proportion of the dealers trade in the inter-dealer market, or a proportion trade with customers, or there is no trading, or the asset matures and a new one is issued. The dealers who trade in each round are chosen randomly and without replacement from the population of dealers. Draws are independent across rounds. Dealers of the same type are always drawn together: if one dealer of type k trades in the inter-dealer (customer-dealer) market, then all dealers of type k trade in the inter-dealer (customer-dealer) market. With probability $\tilde{\pi}_i$, there is inter-dealer trading, but only amongst a proportion $\frac{\pi_i}{\tilde{\pi}_i}$ of the dealers, so that each dealer trades in the inter-dealer market with ex ante probability π_i. Dealer type k buys quantity $q_k^i(\underline{s}_t)$ units of the risky asset from the inter-dealer market at average price per unit $p^i(\underline{s}_t)$. The inter-dealer market is perfectly competitive, so it does not depend on the quantity traded, and there is one price for all dealers, which is why it is not subsripted by k. The justification for assuming a competitive inter-dealer market is discussed in section 2.5.1. With probability $\tilde{\pi}_c$, there is customer-dealer trading, but only amongst a proportion $\frac{\pi_c}{\tilde{\pi}_c}$ of the dealers, so that each dealer trades in the customer-dealer market with ex ante probability π_c. Trade is with customer type j with probability π_j^c, so the probability of trading with any customer can be written as $\pi_c \equiv \sum_j \pi_j^c$. The distinction between customer types is discussed in section 2.3. The customers who trade are drawn uniformly and without replacement from the population of customers. Dealer type k buys quantity $q_k^c(\underline{s}_t)$ units of the risky asset at price per unit $p_k^c(\underline{s}_t)$, where both the price and quantity are negotiated through Nash bargaining. With probability $1 - \tilde{\pi}_i - \tilde{\pi}_c$ there is no trade by any dealer at time t. With probability $1 - \pi_i - \pi_c$ there is no trade by dealer type $k = 1, ..., N_D$ at time t, and dealers just receive interest on their cash and pay inventory costs on the risky asset. At known times T_n $(n = ..., -1, 0, 1, ...)$, the risky asset matures, all dealers receive payoff $V_{T_n}^{\text{issuer}}$ for each unit they hold, and a new risky asset is issued, with dealer type k buying quantity $q_k^{\text{issuer}}(\underline{s}_{T_n})$ at price per unit $p^{\text{issuer}}(s_{T_n})$. The price only depends on the state at the time of issuance, not on any prior states, not on the quantity bought or sold, and the subscript k is dropped since every dealer is offered the same price. This ensures that prices and quantities of inter-dealer and customer-dealer trades only depend on a finite history of states, since the time of issuance, rather than an infinite one. The dealers' problem essentially resets when each risky asset matures.

The quantities $q_k^i(\underline{s}_t)$, $q_k^c(\underline{s}_t)$, and $q_k^{\text{issuer}}(\underline{s}_{T_n})$ represent the signed quantity: positive if the dealer buys, and negative if she sells. For ease of notation I drop the dealers' k subscripts where

we are only considering a single dealer's optmization problem. Henceforth, I denote f_t, p_t^i, q_t^i, p_t^c, q_t^c, p_t^{issuer}, q_t^{issuer} as shorthand for $f(s_t)$, $p^i(\underline{s}_t)$, $q^i(\underline{s}_t)$, $p^c(\underline{s}_t)$, $q^c(\underline{s}_t)$, $p^{\text{issuer}}(s_t)$, $q^{\text{issuer}}(\underline{s}_t)$, respectively.

2.2.1 The dealers' problem

Each dealer maximizes the present value of her cash-flows, and so her value function at time t is given by:

$$V_t^d(\$_t^d, a_t^d, \underline{s}_t) = \max_{\{q_\tau^i, q_\tau^{\text{issuer}}, p_\tau^c, q_\tau^c\}_{\tau=t}^\infty} E_t\left[\sum_{u=0}^{\infty}\delta^u\left(\underbrace{\$_{t+u}^d\left(1-R^{-1}\right)}_{\text{receive interest}} - \underbrace{p_{t+u}q_{t+u}}_{\text{trade}} - \underbrace{f_{t+u}.(a_{t+u}^d)^2}_{\text{inventory cost}} + \underbrace{a_{T_n}^d V_{T_n} 1_{\{t+u\in\{T_n\}_n\}}}_{\text{maturity}}\right)\right] \tag{1}$$

subject to (p_τ^c, q_τ^c) being the solution to Nash bargaining between dealer and customer, and the budget constraints that the dealer could be facing, described below. Here $\delta \in (0,1)$ is the impatience parameter, and p and q are the price and quantity, respectively, of a trade with the dealers, a customer, or the issuer. $\$_\tau^d$ includes accrued interest received by the dealer at time τ. The dealer's pre-interest cash holding was thus $R^{-1}\$_\tau^d$, so the interest received is $\$_\tau^d - R^{-1}\$_\tau^d = \$_\tau^d\left(1-R^{-1}\right)$. The value function can also be written as a function of the other dealers' current inventories and the current state s_t, instead of the history of states $\underline{s}_t$. Each round, the possible situations are: (1) no trade, (2) trade in the inter-dealer market, or with a customer, and (3) maturity and reissuance. The budget constraints are given by:

1. no trade at time τ:

$$\$_{\tau+1}^d = R\left(\$_\tau^d - \underbrace{f_\tau.(a_\tau^d)^2}_{\text{inventory cost}}\right) \tag{2}$$

$$a_{\tau+1}^d = a_\tau^d \tag{3}$$

2. trade in the inter-dealer market or with a customer at time τ:

$$\$_{\tau+1}^d = R\left(\$_\tau^d - \underbrace{f_\tau.(a_\tau^d)^2}_{\text{inventory cost}} - \underbrace{p_\tau q_\tau}_{\text{trade}}\right) \tag{4}$$

$$a_{\tau+1}^d = a_\tau^d + q_\tau \tag{5}$$

where $(p, q) \in \left\{(p^c_\tau, q^c_\tau), (p^i_\tau, q^i_\tau)\right\}$;

3. maturity and reissuance at time τ:

$$\$^d_{\tau+1} = R\left(\$^d_\tau - \underbrace{f_\tau.(a^d_\tau)^2}_{\text{inventory cost}} + \underbrace{a^d_\tau V^{\text{issuer}}_\tau}_{\text{maturity}} - \underbrace{p^{\text{issuer}}_\tau q^{\text{issuer}}_\tau}_{\text{re-issuance}}\right) \tag{6}$$

$$a^d_{\tau+1} = q^{\text{issuer}}_\tau \tag{7}$$

When there is no trade at time τ, the dealer pays the cost for their inventory of risky assets, and receives interest on their cash holding. After receiving the interest and paying the holding cost, the dealer is left with $\$^d_\tau - f_\tau.(a^d_\tau)^2$ in her cash account, which earns gross interest at gross rate R by time $\tau+1$. When there is a trade at time τ, whether with a customer, dealer, or the issuer, the dealer buys q_τ units of the asset at price per unit p_τ. Thus the dealer's cash outflow $p_\tau q_\tau$ is also deducted from the dealer's cash holding at τ. At maturity the dealer receives V^{issuer}_τ for each unit she holds, and can buy a new issue at price per-unit p^{issuer}_τ, so her additional cashflow is $a^d_\tau V^{\text{issuer}}_\tau - p^{\text{issuer}}_\tau q^{\text{issuer}}$.

In the appendix, I show that if $\delta R < 1$, the dealer's value function takes the following general form at time τ:

$$V^d_\tau(\$^d_\tau, a^d_\tau, \underline{s}_\tau) = \gamma^{\$}(\underline{s}_\tau)\$^d_\tau + \gamma^a(\underline{s}_\tau)a^d_\tau + \gamma^{aa}(\underline{s}_\tau).(a^d_\tau)^2 + \gamma(\underline{s}_\tau) \tag{8}$$

for some functions $\gamma^{\$}(\underline{s}_\tau)$, $\gamma^a(\underline{s}_\tau)$, $\gamma^{aa}(\underline{s}_\tau)$, and $\gamma(\underline{s}_\tau)$. I conjecture that the value function takes this form since it is a similar form to that of the dealer's per-period utility function: linear in her cash holding and quadratic in her risky asset position. In the appendix I verify that if the value function takes this form at some point in time, then by backwards induction on time, it takes that form at all prior periods.

2.3 The customers

There are a continuum of risk-neutral customers, of different types. Customer type j is defined by his conditional expected valuation of the asset, $V^{c_j}_t \equiv E^j_t[V_{t+1}]$, bargaining power $\eta_j \in [0, 1]$, and interest rate $R^{c_j}_t \equiv R^{c_j}(s_t)$. They all have discount rate δ^c. They trade at most once, at birth, with a dealer[6], where the price and quantity are determined by Nash bargaining. Because they are risk-neutral, it follows that the price and quantity they negotiate do not depend on their holdings of either the risk-free or risky asset. They consume all their wealth one period later.

[6]Direct trade between 2 customers is extremely rare in the US corporate bond market.

Heterogeneity in customer bargaining power and valuation allows me to capture the empirical fact in the US corporate bond market that at the same point in time different customer-dealer trades may occur at different prices and quantities. There is anecdotal evidence from industry professionals that different customers receive different prices based on their importance to the dealer, with customers anticipated to give a lot of future business to dealers rewarded with better prices. This is why heterogeneity in customer bargaining power is a feature of my model. Given that there are sometimes customer-dealer trades both to buy and sell, at the same time, for the same price, allowing for differences in opinion in customer valuation is also necessary. Without heterogeneity in customer interest rates, an outcome of Nash bargaining would be that customer bargaining power would not affect trade size, so smaller customers with lower bargaining power and tighter budget constraints would trade the same size as larger customers. But if lower customer bargaining power is associated with more adverse interest rates (higher when borrowing, lower when saving), a positive relationship between trade size and customer bargaining power can be established.

2.4 Customer-dealer trade

When a customer and dealer meet they bargain over both the price and quantity, maximizing the Nash product:

$$\max_{p_t^c, q_t^c} [\text{customer gain from trade}]^{\eta} \times [\text{dealer gain from trade}]^{1-\eta} \tag{9}$$

where η is the relative bargaining power of the customer, and $1-\eta$ is that of the dealer. The customer's gain from trade is the utility they get from trading q_t^c units of the risky asset with the dealer at average price p_t^c, over and above the utility from not trading at all. The dealer's gain from trade is their value function if they trade q_t^c units of the risky asset with the customer at price per unit p_t^c, over and above their value function if they did not trade at all.

2.4.1 Customer's gain from trade with a dealer

Consider a customer whose conditional expected valuation of the risky asset is V_t^c per unit, and whose interest rate is R_t^c. For a customer-dealer trade at time t, if a dealer buys q_t^c units of the asset from that customer, at price per unit q_t^c, I show in the appendix that the customer's gain from trade is given by:

$$\text{Customer's gain from trade} = \delta^c R_t^c q_t^c \left(p_t^c - \tilde{V}_t^c \right) \tag{10}$$

where $\tilde{V}_t^c \equiv V_t^c/R_t^c$ is the customer's expected valuation of the asset, discounted back to time t. The gain does not depend on the customer's pre-trade holdings in either cash or the risky asset, because he is risk-neutral. All he cares about is the difference between his discounted valuation and the price at which he can trade with the dealer, i.e. his marginal gain for each unit traded. This is scaled up by the quantity traded, and discounted. Fixing the price per unit, he would want to trade an unbounded amount: buying as much as possible if his valuation is above the dealer's quoted price, and selling otherwise. But the convexity of the dealers' holding cost make this unappealing to the dealer, and bounds the agreed quantity to a finite amount.

2.4.2 Dealer's gain from trade with a customer

Let q_t^c denote the optimal quantity the dealer would buy if she meets a customer at time t. Let q_{t+u+1}^i and q_{t+u+1}^c denote the optimal quantity the dealer would buy if she trades in the inter-dealer market or with a customer at time $t+u+1$, respectively, assuming her last trade was buying q_t^c at time t. If a dealer enters the trade with inventory a_t^d, I show in the appendix that her gain from trade from buying q_t^c units at price per unit $p_t^c(a_t^d, q_t^c)$, versus not trading, can be decomposed as follows:

1. buy q_t^c today at price per unit p_t^c, with a cash-flow of $-p_t^c q_t^c$;
2. pay per-period inventory costs of $f.(a_t^d+q_t^c)^2$ instead of $f.(a_t^d)^2$ every period until the next trade;
3. buy q_{t+u}^i units instead of $q_{t+u}^i + q_t^c$ units at time $t+u$, if the next trade is at time $t+u$ and is in the inter-dealer market;
4. buy q_{t+u}^c units instead of $q_{t+u}^c + q_t^c$ units at time $t+u$, if the next trade is at time $t+u$ and is with a customer;
5. receive V_T^{issuer} per unit on $a_t^d + q_t^c$ units instead of a_t^d units at maturity time T, if there is no trade before time T.

(3) and (4) follow since I will show that the dealer's post-trade inventory position does not depend on her pre-trade inventory. The total gain from trade is the sum each of these 5 pieces, multiplied by the probability they occur, and discounted back to time t. The probability that the dealer doesn't trade is $1-\pi_i-\pi_c$ each period. The probability that the next trade after time t is in the inter-dealer market, and is at time $t+u$ is $(1-\pi_i-\pi_c)^{u-1}\pi_i$. The probability that the next trade after time t is with a customer of type j, and is at time $t+u$ is $(1-\pi_i-\pi_c)^{u-1}\pi_j^c$. The probability that there is no trade until maturity at time T is $(1-\pi_i-\pi_c)^{T-t-1}$. Putting

these pieces together, I show in the appendix that her gain from trade is given by:

$$
\begin{aligned}
(1+\delta R\gamma^{\$})\Bigg(& \underbrace{-p_t^c q_t^c}_{\substack{\text{(1) buy } q_t^c \text{ units from} \\ \text{time-}t \text{ customer}}} \quad \underbrace{-\delta \overbrace{\left(2a_t^d+q_t^c\right)q_t^c}^{(a_t^d+q_t^c)^2-(a_t^d)^2} \sum_{u=0}^{T-t-1}\left(\delta(1-\pi_i-\pi_c)\right)^u E_t[f_{t+u+1}]}_{\substack{\text{(2) discounted inventory costs on } a_t^d+q_t^c \text{ units instead} \\ \text{of } a_t^d, \text{ until next expected trade}}} \\
& +\delta \sum_{u=0}^{T-t-2}\left(\delta(1-\pi_i-\pi_c)\right)^u \left(\pi_i E_t[p_{t+u+1}^i]\left(-q_{t+u+1}^i-(-(q_{t+u+1}^i+q_t^c))\right)\right. \\
& \underbrace{\left.+\sum_j \pi_j^c E_t\left[\underbrace{-q_{t+u+1}^{c_j}\cdot p_{t+u+1}^{c_j}\left(a_t^d+q_t^c,q_{t+u+1}^{c_j}\right)}_{\text{trade at } t \text{ and } t+u+1}-\underbrace{\left(-\left(q_t^c+q_{t+u+1}^{c_j}\right)\cdot p_{t+u+1}^c\left(a_t^d,q_t^c+q_{t+u+1}^{c_j}\right)\right)}_{\text{trade at } t+u+1 \text{ only}}\right]\right)}_{\substack{\text{(3) and (4): Buy } q_{t+u+1} \text{ units from time-}(t+u+1) \text{ counterparty, instead of } q_{t+u+1}+q_t^c. \text{ The} \\ \text{counterparty is a dealer with probability } \pi_i, \text{ and customer } j \text{ with probability } \pi_j^c.}} \\
& + \underbrace{\left(\delta(1-\pi_i-\pi_c)\right)^{T-t-1}\delta\left((a_t^d+q_t^c)-a_t^d\right)E_t\left[V_T^{\text{issuer}}\right]}_{\substack{\text{(5) Cash in } a_t^d \text{ units instead of } (a_t^d+q_t^c) \text{ at maturity time } T, \text{ if dealer} \\ \text{doesn't meet a counterparty in the previous } T-t-1 \text{ rounds of trading.}}}\Bigg)
\end{aligned}
\tag{11}
$$

The sum of these 5 pieces is multiplied by $1+\delta R\gamma_{t+1}^{\$}$. The '1' represents the cash-flow at time t. In the appendix I show that $\gamma_t^{\$}$ is constant, which is why I drop the t subscript in the expression above. $\delta R\gamma_{t+1}^{\$}$ represents the opportunity cost of not investing that one unit of cash-flow in the cash account: it would accrue interest at rate R, be worth $\gamma_{t+1}^{\$}$ next period, but be discounted back at rate δ.

2.4.3 Customer-dealer price

An outcome of Nash bargaining, which I verify in the appendix, is that the customer-dealer price can be expressed as a weighted average of the dealer's and customer's valuations, where the weights are given by their relative bargaining power:

$$p_t^c = \eta V_t^d + (1-\eta)\tilde{V}_t^c \tag{12}$$

where η is the customer's relative bargaining power, $1-\eta$ the dealer's, V_t^d the dealer's time-t valuation of the risky asset, and $\tilde{V}_t^c$ the customer's. The customer's valuation is exogenous; it is their discounted reservation value for the asset. The dealer's valuation is endogenous, and can be characterized similarly as their break-even price; it is the customer-dealer price such that

the dealer's gain from trade is zero. Using the expression for the dealer's gain from trade, the customer-dealer price per unit can thus be written recursively:

$$
\begin{aligned}
p_t^c \;=\; \eta \Bigg(& \underbrace{-\delta\left(2a_t^d + q_t^c\right) \sum_{u=0}^{T-t-1} \left(\delta(1-\pi_i-\pi_c)\right)^u E_t[f_{t+u+1}]}_{\text{inventory costs on } a_t^d+q_t^c \text{ units instead of } a_t^d, \text{ until next trade}} \\
& + \delta \sum_{u=0}^{T-t-2} \left(\delta(1-\pi_i-\pi_c)\right)^u \times \\
& \underbrace{\left(\pi_i E_t[p_{t+u+1}^i] + \sum_j \pi_j^c E_t \left[\underbrace{-q_{t+u+1}^{c_j} . p_{t+u+1}^{c_j}\left(q_{t+u+1}^{c_j}\right)}_{\text{trade at } t \text{ and } t+u+1} - \underbrace{\left(-\left(q_t^c + q_{t+u+1}^{c_j}\right) . p_{t+u+1}^c \left(q_t^c + q_{t+u+1}^{c_j}\right)\right)}_{\text{trade at } t+u+1 \text{ only}} \right] / q_t^c \right)}_{\begin{array}{c}\text{buy } q_{t+u+1} \text{ units from time-}(t+u+1) \text{ counterparty, instead of } q_{t+u+1}+q_t^c. \text{ The counterparty is} \\ \text{a dealer with probability } \pi_i, \text{ and customer } j \text{ with probability } \pi_j^c\end{array}} \\
& + \underbrace{\left(\delta(1-\pi_i-\pi_c)\right)^{T-t-1} \delta\left(\left(a_t^d+q_t^c\right)-a_t^d\right) E_t\left[V_T^{\text{issuer}}\right]}_{\begin{array}{c}\text{cash in } a_t^d \text{ units instead of } (a_t^d+q_t^c) \text{ from issuer at time } T, \text{ if dealer} \\ \text{doesn't meet a counterparty in the previous } T-t-1 \text{ rounds of trading}\end{array}} \Bigg) + (1-\eta)\tilde{V}_t^c
\end{aligned}
\tag{13}
$$

There are two ways that inventory affects the price at which dealers trade with customers: firstly it affects the expected inventory cost until the next trade; secondly it affects the price of the next trade if it is with a customer, as it affects the dealer's bargaining position at that trade.

Theorem 2.1. *The per-unit customer-dealer price, whether the dealer is buying from, or selling to, a customer, can be written as:*

$$
p_t^c \;=\; \tilde{V}_t^c + \eta\delta q_t^c \sum_{u=0}^{T-t-1} \left(\delta\tilde{\pi}\right)^u E_t[f_{t+u+1}] \tag{14}
$$

where $\tilde{\pi} \equiv 1-\pi_i-\sum_j \pi_j^c(1-\eta_j)$.

The proof is given in the appendix. We see that the customer-dealer price can be benchmarked to the customer's discounted valuation, but will usually deviate from it. The direction of the deviation depends solely on whether the customer is buying or selling, i.e. the sign of q_t^c, since the other parameters are positive. If q_t^c is positive, i.e. the dealer is buying from a customer, then the price per unit increases in the customer's bargaining power, dealers' expected inventory costs, and absolute trade size. If q_t^c is negative, i.e. the dealer is selling to a customer, then the price per unit decreases in the customer's bargaining power, dealers' expected inven-

tory costs, and absolute trade size. The larger the trade size, the further the price is from the customer's valuation, as the dealers' total gain is scaled up by a larger quantity, so they are willing to accept a worse price per unit. If the customer's bargaining power is zero, then the dealer gets all the gain from trade, the customer none, and the customer-dealer price is equal to the customer's valuation, V_t^c. The higher the customer's bargaining power is, the more the price moves away from their reservation value, and the greater their gain from trade. The higher the dealer's inventory costs are, the worse the dealer's bargaining position, so the more the price moves away from the customer's reservation value, and the greater the gain the dealer gives to the customer. This expression for $p_t^c - \tilde{V}_t^c$, the difference between the price per unit at which a customer can transact, and their discounted valuation, is one theoretical measure of liquidity. But in this paper I will use the inter-dealer price, rather than the customer's valuation, as the benchmark for the customer-dealer price. This will help to test the model empirically, as the customer valuation is never observable, whereas the inter-dealer price is observed, at least at some points in time.

2.4.4 Customer-dealer quantity

Theorem 2.2. *Each dealer's inventory position after trade with a customer, is given by:*

$$\begin{aligned} a_{t+1}^{d,c} &\equiv a_t^d + q_t^c \qquad (15) \\ &= \frac{\delta E_t\left[\sum_{u=0}^{T-t-2}(\delta\tilde{\pi})^u\left(\pi_i p_{t+u+1}^i + \sum_j \pi_j^c(1-\eta_j)V_{t+u+1}^{c_j}\right) + (\delta\tilde{\pi})^{T-t-1}V_T^{issuer}\right] - \tilde{V}_t^c}{2\delta\sum_{u=0}^{T-t-1}(\delta\tilde{\pi})^u E_t\left[f_{t+u+1}\right]} \end{aligned}$$

The proof is given in the appendix. Since the dealers have identical preferences, they all choose to trade to the same position, regardless of the inventory, a_t^d, they have when they arrive to the trade. From the customer's gain from trade, we know that all he cares about is the difference between his discounted valuation and the price at which he can currently trade with the dealer. Fixing the price per unit, he would want to trade an unbounded amount: buying as much as possible if his valuation is above the dealer's quoted price, and selling otherwise. But the convexity of the dealer's inventory cost bounds the agreed quantity to a finite amount. Fixing prices and valuations, the larger the expected inventory cost, the smaller the dealer's post-trade inventory position. Pre-trade inventory of one trade is post-trade inventory of the previous one. So when inventory costs are higher, both pre- and post-trade inventory will be smaller. Since trade size is the difference between pre- and post-trade inventory, larger inventory costs is associated with smaller trade size. The dealer is not a myopic market-maker who targets

zero inventory to minimize her immediate inventory costs. Instead, fixing inventory costs, she holds a more positive inventory if future expected inter-dealer prices or customer valuations are high relative to the valuation of the customer with whom she is currently trading. Since customer-dealer prices are a weighted average of customer and dealer valuations, high future customer valuations correspond to high future customer-dealer prices, and the dealer can profit from stocking up now relatively cheaply, and selling off later at a higher price. Similarly, if the current customer has a high enough valuation, the dealer sells to him knowing she expects to be able to profit from unwinding the position at the next trade by buying back relatively cheaply. If all the customers have the same interest rates, then a feature inherited from Nash bargaining is that the quantity traded, and thus also the the dealer's post-trade position, do not depend on the customers' bargaining power. If customers with lower bargaining power suffer worse interest rates, i.e. higher when buying and borrowing, and lower when selling and saving, then they will trade in smaller size. In this way smaller customers would trade in smaller size, and then trade size and post-trade inventory would depend on customers' bargaining power.

Figure 6 shows the price-quantity regions where there are gains from trade for the dealer and customer, the equilibrium total price (quantity $\times$ price/unit) and quantity at which the dealer and customer will transact, and how the total gain from trade is split between them, given their relative bargaining powers. The solid and dashed lines represent the indifference curves for the dealer and customer, respectively. The customer's indifference curve is linear since he is risk neutral. The dealer's indifference curve is quadratic due to her inventory costs being quadratic in her risky asset holding position. If there is no trade, i.e. $q_t^c = 0$, then neither the dealer nor customer gains from trade, which is why both indifference curves pass through the origin. There are 4 scenarios depending on the dealer and customer types, which affect the location and slopes of the indifference curves. The region below (above) the solid (dashed) line represents the region where the dealer (customer) has a gain from trade. The intersection of these regions, where there are gains from trade to both dealer and customer, is shaded gray. A feature of Nash bargaining is that the quantity maximizes the total surplus of the customer and dealer. The black dot is plotted at the coordinates of the optimal quantity and total price. The length of the arrows relate to the gains from trade to the customer and dealer. The price is a weighted average of the customer's and dealer's valuations, with the weights given by their relative bargaining power. In this example the customer's relative bargaining power, η, is $\frac{1}{3}$.

2.5 Inter-dealer trade

In the model there is no information asymmetry between dealers, and so risk-sharing drives trading in the inter-dealer market. Reiss and Werner (1998) show that this is the primary motivation for inter-dealer trade in London equity markets. In my model dealers choose to perfectly risk-share in the inter-dealer market: they trade in such a way that their inventory positions after inter-dealer trading are all equal, regardless of the distribution of their inventories before, because they have identical preferences.

Theorem 2.3. *Each dealer's inventory position after trading in the inter-dealer market is given by:*

$$\begin{aligned} a_{t+1}^{d,i} &\equiv a_t^d + q_t^i \\ &= \frac{\delta E_t\left[\sum_{u=0}^{T-t-2}(\delta\tilde{\pi})^u\left(\pi_i p_{t+u+1}^i + \sum_j \pi_j^c(1-\eta_j)V_{t+u+1}^{c_j}\right) + (\delta\tilde{\pi})^{T-t-1}V_T^{issuer}\right] - p_t^i}{2\delta\sum_{u=0}^{T-t-1}(\delta\tilde{\pi})^u E_t\left[f_{t+u+1}\right]} \end{aligned} \tag{16}$$

The proof is given in the appendix. This expression is very similar to the dealer's optimal holding after trade with a customer, but the benchmark which determines the trade direction and size is the current inter-dealer price instead of the current customer valuation. If the inter-dealer price is relatively low, the dealer will stock up on inventory, expecting to sell it in future customer-dealer or inter-dealer markets, or to the issuer, at a higher price. Similarly, if the inter-dealer price is relatively high, the dealer will sell in the inter-dealer market, to buy back relatively cheaply in the future.

Market-clearing in the inter-dealer market implies that net trading amongst the dealers must be zero. This gives a recursion for the inter-dealer price, p_t^i, which can be solved by backwards induction on trade time, starting from the last trade before the maturity of the asset. Let $\bar{a}_t^{d,i}$ denote the mean inventory of dealers who trade in the inter-dealer market at time t. The recursion is given by:

$$\begin{aligned} p_t^i &= \delta E_t\left[\sum_{u=0}^{T-t-2}(\delta\tilde{\pi})^u\left(\pi_i p_{t+u+1}^i + \sum_j \pi_j^c(1-\eta_j)V_{t+u+1}^{c_j}\right) + (\delta\tilde{\pi})^{T-t-1}V_T^{\text{issuer}}\right] \\ &\quad -2\delta\frac{\bar{a}_t^{d,i}}{\pi_i/\tilde{\pi}_i}\sum_{u=0}^{T-t-1}(\delta\tilde{\pi})^u E_t\left[f_{t+u+1}\right] \end{aligned} \tag{17}$$

Given the dynamic nature of the dealer's optimization problem, the current inter-dealer price is increasing in expected future inter-dealer prices, expected future customer valuations, and the expected value of the asset at maturity. The larger the average inventory of dealers trading

in the inter-dealer market, $\bar{a}_t^{d,i}$, is, the greater the supply of the asset, and the lower the price. This is particularly pronounced when dealers' expected inventory costs are high.

Similarly to the customer-dealer price, we can gain intuition for this recursion, by re-writing it. The inter-dealer price per unit is the dealer's gain from trade from buying in the inter-dealer market, differentiated with respective to the number of units bought. If a dealer enters the inter-dealer market with inventory a_t^d, I show in the appendix that the change in her value function from buying a marginal unit at price per unit p_t^i, can be decomposed intuitively as follows:

1. buy q_t^i units today at price per unit p_t^i, with a cash-flow of $-p_t^i q_t^i$;
2. pay per-period inventory costs of $f.(a_t^d + q_t^i)^2$ instead of $f.(a_t^d)^2$ every period until the next trade;
3. buy q_{t+u}^i units instead of $q_{t+u}^i + q_t^i$ units at time $t+u$, if the next trade is at time $t+u$ and is in the inter-dealer market;
4. buy q_{t+u}^c units instead of $q_{t+u}^c + q_t^i$ units at time $t+u$, if the next trade is at time $t+u$ and is with a customer;
5. receive V_T^{issuer} per unit on $a_t^d + q_t^i$ units instead of a_t^d units at maturity time T, if there is no trade before time T.

This means the inter-dealer price can be written as:

$$
\begin{aligned}
p_t^i(\bar{a}_t^{d,i}) \;=\; \frac{\partial}{\partial q_t^i}\Bigg[& \underbrace{-\delta(\bar{a}_t^{d,i})^2 \sum_{u=0}^{T-t-1} (\delta(1-\pi_i-\pi_c))^u E_t[f_{t+u+1}]}_{\text{discounted inventory costs until next expected trade}} + \\
& \sum_{u=0}^{T-t-2} (\delta(1-\pi_i-\pi_c))^u \Big(\pi_i E_t[p_{t+u+1}^i]((q_{t+u+1}^i + q_t^i) - q_{t+u+1}^i) \\
& \quad + \underbrace{\sum_j \pi_j^c ((q_t^i + q_{t+u+1}^c).p_{t+u+1}^{c_j}(a_t^d, q_t^i + q_{t+u+1}^c) - q_{t+u+1}^c.p_{t+u+1}^{c_j}(a_t^d + q_t^i, q_{t+u+1}^c)))}_{\text{sell } q_{t+u+1} \text{ units instead of } q_{t+u+1} + q_t^i \text{ at next trade. Next counterparty is dealer with probability } \pi_i \text{ and customer } j \text{ with probability } \pi^{c_j}.} \\
& + \underbrace{(\delta(1-\pi_i-\pi_c))^{T-t-1} E_t\left[V_T^{\text{issuer}}\right]\left(\left(a_t^d + q_t^i\right) - a_t^d\right)\Big]}_{\text{cash in } a_t^d \text{ units instead of } a_t^d + q_t^i \text{ from issuer at time } T, \text{ if dealer doesn't meet a counterparty in the previous } T-t-1 \text{ rounds of trading}}
\end{aligned}
\tag{18}
$$

2.5.1 Competitiveness of the inter-dealer market

While in reality the inter-dealer market is not perfectly competitive, as in the model, the empirical evidence seems to suggest that it is much more competitive than the customer-dealer market. Figure 5 of Randall (2015) shows the percentage deviation of prices from their mean within a bond issue and within a minute, for customer-dealer and inter-dealer trades, averaged across all paired trades. Price dispersion within a small interval of time is a measure of competitiveness of those prices: the higher the dispersion, the less competitive prices are. By considering trades only within a minute of each other, this is unlikely to represent time series price volatility, but instead the dispersion of prices between different agents at a point in time. Also plotted is the ratio of the price dispersions: customer-dealer versus inter-dealer. This ratio is greater than one for all but one month in the sample, typically around 2, and much higher more recently. That paper also applies a Kalman filter to the unpaired trade prices, which leads to a similar conclusion about customer-dealer versus inter-dealer prices, suggesting that the inter-dealer market is typically much more competitive than the customer-dealer market.

2.6 Issuance

The dealer's inventory position at time T_1, just after the maturity of one risky asset and issuance of a new one, is given by:

$$\begin{aligned} a^d_{T_1+1} &= q^{\text{issuer}}_{T_1} \\ &= \frac{\delta E_{T_1}\left[\sum_{u=0}^{T_2-T_1-2}(\delta\tilde{\pi})^u\left(\pi_i p^i_{T_1+u+1}+\sum_j \pi^c_j(1-\eta_j)V^{cj}_{T_1+u+1}\right)+(\delta\tilde{\pi})^{T_2-T_1-1}V^{\text{issuer}}_{T_2}\right]-p^{\text{issuer}}_{T_1}}{2\delta\sum_{u=0}^{T_2-T_1-1}(\delta\tilde{\pi})^u E_{T_1}\left[f_{T_1+u+1}\right]} \end{aligned} \tag{19}$$

where T_2 is the time of maturity of the new asset which is issued at time T_1. This gives the initial distribution of dealer inventories at the start of the new risky asset's life. Since the issuance price depends only on the current state, and not any of the previous ones, the amount the dealers trade at issuance depends only on prices, valuations, and inventory costs during the life of that risky asset and not any of those objects for past or future issuances: i.e. maturity of the asset resets the dealers' problem. We see that

the dealers' optimal inventory position is very similar to that after customer-dealer or inter-dealer trading, but the benchmark price is now the new issuance price, $p_{T_1}^{\text{issuer}}$, rather than the customer valuation or inter-dealer price.

2.7 Dealer markup

I define the dealer's per-unit markup when trading with a customer to be the difference between the customer-dealer price and expected inter-dealer price, specifically:

$$\text{markup}_t \equiv \begin{cases} E[p_t^i] - p_t^c & \text{if the dealer buys from a customer at time } t \\ p_t^c - E[p_t^i] & \text{if the dealer sells to a customer at time } t \end{cases}$$

If transactions only happen at quoted prices, then the bid-ask spread would be $p_t^{c,ask} - p_t^{c,bid}$, where $p_t^{c,ask}$ is the ask price offered when the customer is buying from the dealer, and $p_t^{c,bid}$ is the bid price when the customer is selling to her. So the markup splits the bid-ask spread into two, possibly unequal, pieces. The benchmark price is not the arbitrary bid-ask midpoint, but the inter-dealer price. We can interpret the markup as the mark-to-market profit of the dealers. It is the analytical analogue to the "unique round trip" measure from Feldhutter (2012), which focuses on paired trades, when there are customer-dealer and inter-dealer trades within a few minutes. In my model only one of these types of trade can occur at each point in time, i.e. I am focusing on the unpaired trades, when dealers have to bear some change in their inventory costs for a non-trivial period of time. But by allowing the possibility of either customer-dealer or inter-dealer trading at each point in time in my model, I can compute the theoretical inter-dealer price at the time of a customer-dealer trade. In the appendix I show that

the markup when the dealer is buying from a customer is given by:

$$E[p_t^i] - p_t^c = \delta \left(\left(\underbrace{2a_t^{d,c} + q_t^c}_{\text{dealer's inventory}} - 2. \underbrace{(a_t^{d,i} + q_t^i)}_{\text{average inventory of dealers in}} \right) \underbrace{}_{\text{cost gain per unit}} \underbrace{}_{\text{hypothetical inter-dealer market}} \right) + (1-\eta) q_t^c \Bigg) \times \sum_{u=0}^{T-t-1} (\delta\tilde{\pi})^u \, E_t[f_{t+u+1}] \tag{20}$$

The expression for the markup when the dealer is selling is the negative of this expression. Whether the dealer is buying or selling, customers with higher bargaining power, η, get lower markups. When a dealer has inventory above the inter-dealer mean, the markup when she is buying is increasing in expected inventory costs, and the markup when she is selling is decreasing in those costs. The greater her inventory is above the mean, the stronger this effect. Thus the model can generate the markup asymmetry for buys versus sells we see during the financial crisis in Figure 2, as long as the dealers trading with customers were mostly those with above-average inventory. Given their increased incentive to offload inventory at that time, it is reasonable to believe that they were the more active dealers.

Whether the dealer trades with a customer or dealer at time t, at time $t+1$ her optimization problem looks the same, though she may enter time $t+1$ with different cash and risky asset positions. So there is a lot of cancelation between customer-dealer and inter-dealer prices at time t, and from the expression for the markup, we see that the residual which doesn't cancel is comprised of three pieces. One piece is proportional to $a_t^{d,i} + q_t^i$, and comes from the expression for the inter-dealer price. Since net trading in the inter-dealer market is zero, all dealers trade to the same position in the inter-dealer market, $\bar{a}_t^{d,i} \equiv a_t^{d,i} + q_t^i$. So from equation (17) we see that the inter-dealer price, and therefore the markup when the dealer is buying, is decreasing in the average inventory of the dealers trading. The other two pieces come from the expression for customer-dealer price. If there's no trade in the time interval $(t,\, t+u+1)$, then the difference in time-$(t+u+1)$ inventory costs of buying q_t^c units of the risky assets from the customer

versus not trading is $((a_t^{d,c}+q_t^c)^2-(a_t^{d,c})^2)f_{t+u+1}=(2a_t^{d,c}+q_t^c)q_t^c f_{t+u+1}$. Dividing this by q_t^c gives the per-unit inventory cost gain, $(2a_t^{d,c}+q_t^c)f_{t+u+1}$. The remaining piece, proportional to $(1-\eta)q_t^c$ shows that the markup is higher, whether a buy or sell, when customer's bargaining power, η, is lower since the customer-dealer price is then less favorable to the customer, as we see from equation (14).

2.8 Finitely-lived dealers

In the previous subsections I assumed that the dealers are infinitely-lived. I also solve a model where the dealers live only until the maturity of the asset. For the sake of brevity, the proofs are omitted, but are very similar to the infinitely-lived case given in the appendix. The expressions for prices and quantities are very similar to those when dealers are infinitely-lived. The only difference is the coefficient on the dealer's cash position in the dealer's value function, $\gamma^{\$}(s_t)$. When dealers are infinitely lived, $\gamma^{\$,inf}(s_t)=\frac{1-R^{-1}}{1-\delta R}$, and when dealers are finitely lived, $\gamma^{\$,fin}(s_t)=\frac{(1-R^{-1})(1-(\delta R)^{T-t+1})}{1-\delta R}$. Figure 7 shows how this coefficient evolves in both cases, where dealers are finitely and infinitely lived. In that example I look at periods up to 1,000 units of time before the maturity of the bond, dealers earn interest at 5% per year on their cash position, and the discount factor, δ, is 0.99 per year. When dealers are infinitely lived, this coefficient is constant: a dollar held by the dealers today or a dollar held by them tomorrow will produce the same infinite stream of interest payments going forward. However, when dealers are finitely-lived, though $\gamma^{\$}(s_t)$ remains deterministic, it is no longer constant, and decreases monotonically as the dealers' death approaches: a dollar today will earn more interest payments than a dollar tomorrow. Consequently, this non-constant value to cash holdings is reflected in prices and quantities. Earlier cash-flows are weighted more heavily than in the infinitely-lived case. Dealers' post-trade inventories are determined by a trade-off between the current price and expected future ones, as they manage their portfolios intertemporally. Their positions weigh more heavily on the current customer valuation for customer-dealer trades and on the current inter-dealer price for inter-dealer trades, for finitely-lived dealers than infinitely-lived ones. The further the trade is from the dealers' death, the closer the expressions for prices and quantities for the finitely-lived case are to the infinitely-lived ones. The dealers' post-trade holdings, when they are finitely lived, are given by:

$$a_{t+1}^d=a_t^d+q_t= \tag{21}$$
$$\frac{\delta E_t[\sum_{u=0}^{T-t-2}(\delta\tilde{\pi})^u(1+\delta R\gamma_{t+u+1}^{\$})(\pi_i p_{t+u+1}^i+\sum_j \pi_j^c(1-\eta_j)V_{t+u+1}^{cj})+(\delta\tilde{\pi})^{T-t-1}V_T^{\text{issuer}}]-V_t(1+\delta R\gamma_{t+1}^{\$})}{2\delta\sum_{u=0}^{T-t-1}(\delta\tilde{\pi})^u(1+\delta R\gamma_{t+u+1}^{\$})E_t[f_{t+u+1}]}$$

where V_t denotes the customer's discounted valuation $\tilde{V}_t^c$ for a customer-dealer trade, and the inter-dealer price p_t^i for an inter-dealer trade. The markup when a dealer is buying becomes:

$$\begin{aligned}\text{markup}_t^{\text{buy}} = E_t[p_t^i] - p_t^c &= \delta\left(2a_t^{d,c} - 2(a_t^{d,i} + q_t^i) + (2-\eta)q_t^c\right) \\ &\quad \times \sum_{u=0}^{T-t-1} (\delta\tilde{\pi})^u \frac{1+\delta R\gamma_{t+u+1}^{\$}}{1+\delta R\gamma_{t+1}^{\$}} E_t[f_{t+u+1}] \quad (22)\end{aligned}$$

We see that this expression is very similar to the case when dealers are infinitely lived. The only difference is that more weight is put on expected inventory costs in the near future than far future.

3 Empirical Predictions and Evidence

The structural relationships derived earlier allow me to make precise empirical predictions for dealer markups and inventory positions. The model predicts that:

(H1) The higher the dealers' expected inventory costs, the lower the volume they buy and the more they sell, particularly for more risky assets.

(H2) The higher the level of dealer inventory, and the cost of that inventory, the greater the markup will be when the dealer is buying, and the lower the markup when the dealer is selling. These effects should be stronger when expected inventory costs are higher; either because the expected time to trade is longer, or the per-period cost is higher.

(H3) The higher the dealers' expected inventory costs, the lower will be trade size.

(H4) Fixing bargaining power, the markup is increasing in the product of trade size and expected inventory costs, for both buys and sells.

(H5) The higher the customer's bargaining power is, the less the markup is increasing in that product, for both buys and sells.

Figures 1 - 4 show the aggregate stylized time-series patterns for the US corporate bond market in the last several years. Figures 2 - 4 are constructed using an enhanced version of TRACE, a data-set of US corporate bond transactions, cleaned using the methodology of Dick-Nielsen (2014). Markups are computed using the methodology of Randall (2015). Figure 1 supports hypothesis (H1) by showing how dealers reduced their overall inventory of corporate securities in the financial crisis of 2007-09 using primary dealer inventory data from the Federal Reserve Bank of New York. Choi and Shachar (2014) and "Digging into dealer inventories" (September 11th 2013, FT.com) suggest that the extent of the deleveraging in the corporate bond market may be lower, since the New York Fed data includes corporate securities other than corporate bonds, but all these studies agree dealer inventories are much lower than their pre-crisis peak. Figure 2 supports hypothesis (H2) by showing that in the financial crisis, when dealers had relatively high inventory and were selling, for unpaired customer-dealer trades the median markup when the dealer was buying increased, and the median markup when the dealer was selling decreased. Figure 3 supports hypothesis (H3) by showing that the mean trade size has decreased since the start of the financial crisis, in the inter-dealer market and when dealers were both buying from, and selling to. customers. Figure 4 relates to hypotheses (H4) and (H5), showing the relationship between the markup and trade size. For large trades, in excess of $1 million notional value, where customers are large institutions and so bargaining power is likely to be similar, markups increase in trade size. For small and medium sized trades, markups tend to decrease with trade size, with larger trades associated with higher bargaining power.

In Randall (2015) predictions (H1), (H4), and (H5) are tested in the cross-section of the US corporate bond market in 2004 - 2010. Prediction (H2) is not tested due to lack of data on individual dealer inventory. Credit ratings are used as a measure of inventory costs for individual bonds. Two measures for aggregate dealers' inventory costs are used: the LIBOR-OIS spread which measures dealers' funding liquidity, and the ratio of conditional equity return variance for dealers versus customers, which proxies for their relative ability and appetite to bear inventory risk. Bargaining power is proxied by splitting into trade size buckets, with larger trades more likely to be transacted by

more important customers with more bargaining power, and then the effect of trade size is analyzed within those buckets. There is strong statistical support for predictions (H1), (H4), and (H5). A one standard deviation increase in dealers' inventory costs was associated with: an increase of 41% in net selling of high yield bonds and a decrease in net selling of 38% of investment grade ones; a 7.1% increase in markups for unpaired trades in high yield bonds, and a 5.7% increase for investment grade bonds, for a trade of $100,000. For paired trades, where inventory risk should not matter, the effects from inventory costs were much less significant. In the financial crisis inventory costs were dramatically higher than average, so these effects were much larger.

4 Conclusion

My structural model for trading in an OTC search market allows me to show how dealers' inventory costs impact liquidity provision and dealers' optimal inventory positions. It rationalizes the most prominent stylized time-series facts for the US corporate bond market in recent years, with dealers' inventory levels, trade size, and transaction cost asymmetry (buys versus sells) all correlated with dealers' time-varying aversion to retaining inventory for significant periods of time. In addition, Randall (2015) shows that my model correctly predicts that, controlling for customer bargaining power, when inventory costs are high enough, the generally negative relationship between dealer markups and trade size can be reversed, and realigned with the relationship in equity markets. As predicted, this effect is stronger for both riskier bonds and trades associated with customers with lower bargaining power. Without a structural equilibrium model, where trade size is endogenous, customers and dealers bargain over both price and quantity, dealers bear a time-varying aversion to holding inventory, and dealers are long-lived, these relationships would be harder to identify empirically and to interpret.

The US corporate bond market was a natural place for the predictions of my model to be tested empirically, given that my modeling choices were informed by the stylized time-series facts of that market, and the availability of a comprehensive set of transaction level data. But my model can also be tested in other OTC markets. The

Financial Industry Regulatory Authority (FINRA) has also constructed a database of transactions for structured product markets: Asset-Backed Securities, Collateralized Mortgage Obligations, Mortgage-Backed Securities and To Be Announced Securities, whose liquidity has been studied by Friewald, Jankowitsch, and Subrahmanyam (2012). Municipal bond and swaps markets are other natural testing grounds.

Understanding the effect of inventory inventory costs is particularly relevant in the current regulatory environment. The Volcker Rule, part of the Dodd-Frank Act, will restrict commercial banks from proprietary trading. Many of the major corporate bond dealers are subsidiaries of commercial banks, and will thus be affected. The Financial Stability Oversight Council has proposed that proprietary trading may be identified by the regulators as monitoring (a) the ratio of daily turnover of assets to the banks' inventory; and (b) the ratio of customer-initiated trades to dealer-initiated trades, relative to their past trading behavior, other trading desks within the industry, and hedge funds. Banks which hold positions which are too large for too long are more likely to be cited for speculation, and would find it harder to argue these positions were purely for market-making. So going forward a component of dealers' cost of holding inventory is likely to be the risk of breaching this and other regulations. Though these may be effective in limiting opportunities for the dealing arms of banks to speculate, and thus protect depositors, they also restrict dealers' ability to make a market in infrequently traded corporate bonds and other illiquid markets. Understanding this mechanism, and given the feedback from reduced secondary market liquidity to increased cost of capital in the primary market, informs the policy debate.

References

[1] Acharya V. and M. Richardson (editors), Restoring Financial Stability: How to Repair a Failed System, 2009. New York University Stern School of Business

[2] Adrian T., Etula E., and H. Shin, 2010. Risk Appetite and Exchange Rates. Working paper, Federal Reserve Bank of New York

[3] Afonso G., and R. Lagos, 2015. Trade Dynamics in the Market for Federal Funds. *Econometrica*, 83, 1, 263313

[4] Amihud, Y. and H. Mendelson, 1980. Dealership Markets: Market-making with Inventory. *Journal of Financial Economics* 8, 31-53

[5] Choi, J. and O. Shachar, 2014. Did Liquidity Providers Become Liquidity Seekers? Evidence from the CDS-Bond Basis During the 2008 Financial Crisis. Working Paper, University of Illinois Urbana-Champaign & New York Federal Reserve

[6] Dick-Nielsen, J. 2014. How to Clean Enhanced TRACE Data. Working Paper, Copenhagen Business School

[7] Dick-Nielsen, J., P. Feldhutter, and D. Lando, 2012. Corporate Bond Liquidity Before and After the Onset of the Subprime Crisis. *Journal of Financial Economics* 103, 471-492.

[8] Duffie D., N. Garleanu and L. Pedersen, 2005. Over-The-Counter Markets. *Econometrica* 73, 1815-1847.

[9] Duffie D., N. Garleanu and L. Pedersen, 2007. Valuation in Over-The-Counter Markets. *The Review of Financial Studies* 20, 1865-1900.

[10] Etula E., 2013. Broker-Dealer Risk Appetite and Commodity Returns. *Journal of Financial Econometrics* 11, 486-521

[11] Feldhutter P., The same bond at different prices: identifying search frictions and selling pressures. 2012. *Review of Financial Studies*, 25, 1155-1206.

[12] Friewald N., R. Jankowitsch, and M. Subrahmanyam. Liquidity, Transparency and Liquidity in the Securitized Product Market. 2014. Working paper, Vienna University of Economics and Business and New York University.

[13] Garleanu N., Portfolio Choice and pricing in illiquid markets. 2008. *Journal of Economic Theory* 144, 532-564

[14] Lagos R., and G. Rocheteau, Liquidity in Asset Markets with Search Frictions. 2009. *Econometrica* 77, 2, 403-426

[15] Lagos R., G. Rocheteau, and P-O Weill, Crises and Liquidity in Over-the-Counter Markets. 2011. *Journal of Economic Theory* 146, 2169-2205

[16] Longstaff, F., Optimal Portfolio Choice and the Valuation of Illiquid Securities. 2001. *Review of Financial Studies* 14 (2) 407-431.

[17] Longstaff, F., Portfolio Claustrophobia: Asset Pricing in Markets with Illiquid Assets. 2009. *American Economic Review* 99, 4, 1119-1144

[18] Randall, O., How do Inventory Costs affect Dealer Behavior in the US Corporate Bond Market? 2015. Working Paper, Emory University

[19] Reiss, P. and I. Werner, Does Risk Sharing Motivate Interdealer Trading? 1998. *Journal of Finance* 53, 5.

[20] Stoll, H., The Supply of Dealer Services in Securities Markets. 1978. *Journal of Finance* 33, 1133-51.

[21] Weill, P. Leaning against the Wind. 2007. *Review of Economic Studies* 74, 1329-1354.

[22] Zitzewitz, E., Paired Corporate Bond Trades. 2010. Working Paper, Dartmouth College

Appendix A Proofs

Appendix A.1 Value function conjecture

Assume the value function takes the following form, at each time t:

$$V^d_{t+1}(\$^d_{t+1}, a^d_{t+1}, s_{t+1}) \quad = \quad \gamma^{\$}_{t+1}\$^d_{t+1} + \gamma^a_{t+1}a^d_{t+1} + \gamma^{aa}_{t+1}(a^d_{t+1})^2 + \gamma_{t+1}s_{t+1} \tag{1}$$

where $\gamma^{\$}_t \equiv \gamma^{\$}(s_t)$, $\gamma^a_t \equiv \gamma^a(s_t)$, $\gamma^{aa}_t \equiv \gamma^{aa}(s_t)$, and $\gamma_t \equiv \gamma(s_t)$. We proceed by backwards induction on trade time, starting at the time of maturity and re-issuance of the risky asset.

Appendix A.1.1 Maturity and re-issuance

At time T the asset matures, and all dealers receive payoff V^{issuer}_T from the issuer for each unit they hold. The asset is reissued, and dealers can buy unrestricted quantity q^{issuer}_T at price per unit p^{issuer}_T, which depends only on the state at time T. She arrives at time T with a^d_T units of the risky asset, and $\$^d_T$ in cash which includes interest accumulated for the period $(T-1, T]$, which is paid at T. At time T her value function is given by:

$$\begin{aligned} V^d_T(\$^d_T, a^d_T, s_T) \quad = \quad \max_{q^{\text{issuer}}_T} \Bigg\{ & \underbrace{\$^d_T(1-R^{-1})}_{\text{receive interest}} - \underbrace{f_T(a^d_T)^2}_{\text{inventory cost}} + \underbrace{a^d_T V^{\text{issuer}}_T}_{\text{maturity}} - \underbrace{p^{\text{issuer}}_T q^{\text{issuer}}_T}_{\text{re-issuance}} \\ & + \delta E_T \left[V^d_{T+1} \left(R \left(\$^d_T - \underbrace{f_T(a^d_T)^2}_{\text{inventory cost}} + \underbrace{a^d_T V^{\text{issuer}}_T}_{\text{maturity}} - \underbrace{p^{\text{issuer}}_T q^{\text{issuer}}_T}_{\text{re-issuance}} \right), q^{\text{issuer}}_T, s_{T+1} \right) \right] \Bigg\} \end{aligned} \tag{2}$$

The first order condition with respect to q_T^{issuer} yields the dealer's holding at time T:

$$a_{T+1}^d = q_T^{\text{issuer}} = \frac{p_T^{\text{issuer}}\left(1+\delta R E_T[\gamma_{T+1}^{\$}]\right) - \delta E_T\left[\gamma_{T+1}^{a}\right]}{2\delta E_T\left[\gamma_{T+1}^{aa}\right]} \tag{3}$$

The second order condition with respect to q_T^{issuer} confirms this is a maximum as long as $E_T\left[\gamma_{T+1}^{aa}\right] < 0$. I will later show that it is. Plugging this optimal quantity back in to the value function:

$$\begin{aligned} V_T^d(\$_T^d, a_T^d, s_T) \;=\; & \underbrace{\left(1 - R^{-1} + \delta R E_T\left[\gamma_{T+1}^{\$}\right]\right)}_{\gamma_T^{\$}} \$_T^d + \underbrace{V_T^{\text{issuer}}\left(1+\delta R E_T[\gamma_{T+1}^{\$}]\right)}_{\gamma_T^{a}} a_T^d \underbrace{- \left(1+\delta R E_T[\gamma_{T+1}^{\$}]\right) f_T}_{\gamma_T^{aa}} (a_T^d)^2 \\ & + \underbrace{q_T^{\text{issuer}}\left(\delta q_T^{\text{issuer}} E_T\left[\gamma_{T+1}^{aa}\right] + \underbrace{\delta E_T\left[\gamma_{T+1}^{a}\right] - \left(1+\delta R E_T\left[\gamma_{T+1}^{\$}\right]\right) p_T^{\text{issuer}}}_{-2\delta q_T^{\text{issuer}} E_T[\gamma_{T+1}^{aa}]}\right) + \delta E_T\left[\gamma_{T+1}\right]}_{\gamma_T = \delta\left(E_T[\gamma_{T+1}] - \left(q_T^{\text{issuer}}\right)^2 E_T[\gamma_{T+1}^{aa}]\right)} \end{aligned} \tag{4}$$

Appendix A.1.2 Inter-dealer trade

At each trade time $t = ..., -1, 0, 1, ...$ some dealers can access the inter-dealer market, buying quantity q_t^i at price p_t^i. The inter-dealer market is assumed to be perfectly competitive, and so the dealer takes the price as given. She arrives at time t with a_t^d units of the risky asset, and $\$_t^d$ in cash which includes interest accumulated for the period $(t-1, t]$, which is paid at t. At time t her value function is given by:

$$V_t^d(\$_t^d, a_t^d, s_t) \;=\; \max_{q_t^i}\left\{ \underbrace{\$_t^d(1-R^{-1})}_{\text{receive interest}} - \underbrace{f_t(a_t^d)^2}_{\text{inventory cost}} - \underbrace{p_t^i q_t^i}_{\text{inter-dealer trade}} + \delta E_t\left[V_{t+1}^d\left(R\left(\$_t^d - \underbrace{f_t(a_t^d)^2}_{\text{inventory cost}} - \underbrace{p_t^i q_t^i}_{\text{trade}} \right), a_t^d + q_t^i, s_{t+1} \right) \right] \right\} \tag{5}$$

The first order condition with respect to q_t^i yields the dealer's holding after inter-dealer trading at time t:

$$a_{t+1}^{d,i} \equiv a_t^d + q_t^i = \frac{p_t^i \left(1 + \delta R E_t[\gamma_{t+1}^{\$}]\right) - \delta E_t\left[\gamma_{t+1}^{a}\right]}{2\delta E_t^d\left[\gamma_{t+1}^{aa}\right]} \tag{6}$$

The second order condition with respect to q_t^i confirms this is a maximum as long as $E_t^d\left[\gamma_{t+1}^{aa}\right] < 0$. I will later show that it is. Plugging this optimal quantity back in to the value function:

$$\begin{aligned}
& V_t^d(\$_t^d, a_t^d, s_t) \\
= \; & \underbrace{\$_t^d(1 - R^{-1})}_{\text{interest}} - \underbrace{f_t(a_t^d)^2}_{\text{inventory cost}} - \underbrace{p_t^i(a_{t+1}^d - a_t^d)}_{\text{trade}} + \delta E_t\left[R\left(\$_t^d - \underbrace{f_t(a_t^d)^2}_{\text{inventory cost}} - \underbrace{p_t^i(a_{t+1}^d - a_t^d)}_{\text{trade}}\right)\gamma_{t+1}^{\$} + a_{t+1}^d\gamma_{t+1}^a + (a_{t+1}^d)^2\gamma_{t+1}^{aa} + \gamma_{t+1}\right] \qquad (7) \\
= \; & \underbrace{\left(1 - R^{-1} + \delta R E_t\left[\gamma_{t+1}^{\$}\right]\right)}_{\gamma_t^{\$,i}}\$_t^d + \underbrace{p_t^i(1 + \delta R E_t[\gamma_{t+1}^{\$}])}_{\gamma_t^{a,i}} a_t^d \underbrace{-(1 + \delta R E_t[\gamma_{t+1}^{\$}])f_t}_{\gamma_t^{aa,i}}(a_t^d)^2 \\
& \underbrace{+ a_{t+1}^d\left(\delta a_{t+1}^d E_t\left[\gamma_{t+1}^{aa}\right] + \underbrace{\delta E_t\left[\gamma_{t+1}^{a}\right] - (1 + \delta R E_t\left[\gamma_{t+1}^{\$}\right])p_t^i}_{-2\delta a_{t+1}^d E_t[\gamma_{t+1}^{aa}]}\right) + \delta E_t\left[\gamma_{t+1}\right]}_{\gamma_t^i = \delta\left(E_t[\gamma_{t+1}] - (a_{t+1}^d)^2 E_t[\gamma_{t+1}^{aa}]\right)} \qquad (8)
\end{aligned}$$

Appendix A.1.3 Customer-dealer trade

The dealer trades with customer j at each trade time with probability π_j^c. We can think of this probability as the joint probability of the dealer and customer meeting, and there being a gain from trade.

The dealer's gain from trade

Suppose a dealer meets a customer at time t. The dealer's value function[7] if she buys optimal quantity q_t^c at average price per unit p_t^c is given by:

$$\tilde{V}_t^d\left(\$_t^d, a_t^d, s_t, q_t^c\right) = \underbrace{\$_t^d(1-R^{-1})}_{\text{receive interest}} - \underbrace{f_t(a_t^d)^2}_{\text{inventory cost}} - \underbrace{p_t^c q_t^c}_{\text{customer trade}} + \delta E_t\left[V_{t+1}^d\left(R\left(\$_t^d - \underbrace{f_t(a_t^d)^2}_{\text{inventory cost}} - \underbrace{p_t^c q_t^c}_{\text{trade}}\right), a_t^d + q_t^c, s_{t+1}\right)\right] \tag{9}$$

The gain from the dealer buying quantity q_t^c at price p_t^c as opposed to not trading (i.e. $q_t^c = 0$) is given by:

$$\begin{aligned} \text{Dealer's gain from trade} &= \tilde{V}_t^d\left(\$_t^d, a_t^d, s_t, q_t^c\right) - \tilde{V}_t^d\left(\$_t^d, a_t^d, s_t, 0\right) &(10)\\ &= q_t^c E_t\left[\delta\gamma_{t+1}^a + \delta\left(2a_t^d + q_t^c\right)\gamma_{t+1}^{aa} - \left(1+\delta R\gamma_{t+1}^{\$}\right)p_t^c\right] &(11)\end{aligned}$$

The customer's gain from trade

There are a continuum of customers who live for one unit of time, and trade only at birth. Consider one customer with an endowment of ($cash, risky asset units) $= (\$_t^c, a_t^c)$ entering trade at time t. He is risk-neutral, with utility over post-trade wealth W_{t+1} given by $u^c(W_{t+1}) = \delta^c E_t[W_{t+1}] = \delta^c\left(\$_{t+1}^c + a_{t+1}^c V_t^c\right)$, where V_t^c is the expected payoff to the customer of one unit of the risky asset. His interest rate is R_t^c. His gain from selling q_t^c units

[7] I denote it $\tilde{V}_t^d(\$, a, s, q)$ since V_t^d is a function of $(\$, a, s)$, but $\tilde{V}_t^d(\$, a, s, q) \equiv V_t^d(\$, a, s)$ if q is optimal.

of the risky asset to the dealer at an average price p_t^c per unit is given by:

$$\begin{aligned} \text{Customer's gain from trade} &= \underbrace{\delta^c \left(R_t^c \left(\cancel{\$_t^c} + p_t^c q_t^c\right) + (\cancel{a_t^c} - q_t^c) V_t^c\right)}_{\text{utility if trade}} - \underbrace{\delta^c \left(\cancel{R_t^c \$_t^c} + \cancel{a_t^c V_t^c}\right)}_{\text{utility if no trade}} \quad (12) \\ &= \delta^c R_t^c q_t^c \left(p_t^c - \tilde{V}_t^c\right) \quad (13) \end{aligned}$$

where $\tilde{V}_t^c \equiv V_t^c / R_t^c$. Note that the customer's gain from trade doesn't depend on his pre-trade holdings in either cash or the risky asset, because he is risk-neutral.

For both the customer and dealer to have strictly positive gains from trade, we need p_t^c and q_t^c to exist such that:

$$q_t^c \tilde{V}_t^c < q_t^c p_t^c < \frac{\delta q_t^c E_t \left[\gamma_{t+1}^{aa} \left(\frac{\gamma_{t+1}^a}{\gamma_{t+1}^{aa}} + 2a_t^d + q_t^c\right)\right]}{1 + \delta R E_t[\gamma_{t+1}^{\$}]} \quad (14)$$

The RHS is a negative quadratic in q_t^c, with roots at $q_t^c = 0$ and $q_t^c = -E_t\left[\frac{\gamma_{t+1}^a}{\gamma_{t+1}^{aa}} + 2a_t^d\right]$, and whose slope at $q_t^c = 0$ is $\frac{\delta E_t\left[\gamma_{t+1}^a + 2a_t^d \gamma_{t+1}^{aa}\right]}{1+\delta R E_t[\gamma_{t+1}^{\$}]}$. The LHS is linear in q_t^c, with positive slope $\tilde{V}_t^c$, also with its root at $q_t^c = 0$. Thus there are gains from trade as long as $\frac{\delta E_t\left[\gamma_{t+1}^a + 2a_t^d \gamma_{t+1}^{aa}\right]}{1+\delta R E_t[\gamma_{t+1}^{\$}]} \neq \tilde{V}_t^c$. Figure 6 shows the regions where there are gains from trade, the equilibrium total price and quantity at which the dealer and customer will transact, and how the total gain from trade is split between them, given their relative bargaining power. We now derive the price and quantity at which they trade.

The bargaining game between the dealer and the customer

Let $\eta \in [0,1]$ and $1-\eta$ be the relative bargaining powers of the customer and dealer, respectively. They bargain over both price and quantity to maximize the Nash product:

$$\max_{p_t^c, q_t^c} \left\{ \delta^c R^c q_t^c \left(p_t^c - \tilde{V}_t^c \right)^{\eta} \left(E_t \left[\delta \gamma_{t+1}^a + \delta \left(2a_t^d + q_t^c \right) \gamma_{t+1}^{aa} - \left(1 + \delta R \gamma_{t+1}^{\$} \right) p_t^c \right] \right)^{1-\eta} \right\} \tag{15}$$

The first-order condition (FOC) wrt q_t^c yields:

$$p_t^c = \frac{\delta E_t \left[\gamma_{t+1}^a + \left(2a_t^d + (2-\eta) q_t^c \right) \gamma_{t+1}^{aa} \right]}{1 + \delta R E_t [\gamma_{t+1}^{\$}]} \tag{16}$$

The FOC wrt p_t^c yields:

$$\Rightarrow p_t^c = \eta \left(\frac{\delta E_t \left[\gamma_{t+1}^a + \left(2a_t^d + q_t^c \right) \gamma_{t+1}^{aa} \right]}{1 + \delta R E_t [\gamma_{t+1}^{\$}]} \right) + (1-\eta) \tilde{V}_t^c \tag{17}$$

Notice that the customer-dealer price is a weighted average of the dealer's and customer's valuations, i.e. $p_t^c = \eta V_t^d + (1-\eta)\tilde{V}_t^c$, where V_t^d is the price when the dealer's gain is zero. Combining the FOCs (16) and (17), yields the dealer's position after she trades with a customer at time t:

$$a_{t+1}^{d,c} \equiv a_t^d + q_t^c = \frac{\left(1 + \delta R E_t [\gamma_{t+1}^{\$}] \right) \tilde{V}_t^c - \delta E_t \left[\gamma_{t+1}^a \right]}{2\delta E_t \left[\gamma_{t+1}^{aa} \right]} \tag{18}$$

Notice that the dealer's new position after trading with the customer does not depend on her inventory before.

The dealer's value function at time t, before she knows if she will trade with a customer, is given by:

$$V_t^{d,c_j}\left(\$_t^d, a_t^d, s_t\right) = \underbrace{\left(1 - R^{-1} + \delta R E_t[\gamma_{t+1}^{\$}]\right)}_{\gamma_t^{\$,c_j}} \$_t^d + \underbrace{\left(\left(1 + \delta R E_t[\gamma_{t+1}^{\$}]\right)(1-\eta_j) V_t^{c_j} + \delta \eta_j E_t[\gamma_{t+1}^{a}]\right)}_{\gamma_t^{a,c_j}} a_t^d \tag{19}$$
$$+ \underbrace{\left(\delta \eta_j E_t[\gamma_{t+1}^{aa}] - \left(1 + \delta R E_t[\gamma_{t+1}^{\$}]\right) f_t\right)}_{\gamma_t^{aa,c_j}} (a_t^d)^2 + \underbrace{E_t\left[\delta \gamma_{t+1} - a_{t+1}^{d,c_j}\left(\delta \eta_j \left(\gamma_{t+1}^{a} + \gamma_{t+1}^{aa} a_{t+1}^{d,c_j}\right) + (1-\eta_j)\left(1 + \delta R \gamma_{t+1}^{\$}\right) V_t^{c_j}\right)\right]}_{\gamma_t^{c_j}}$$

Appendix A.1.4 Periods of no trading

In periods when no dealer trades, neither with a customer, another dealer, or the issuer, she still gets paid interest, and pays the holding cost. Her value function is given by:

$$V_t^d(\$_t^d, a_t^d, s_t) = \underbrace{(1 - R^{-1} + \delta R E_t^d[\gamma_t^{\$}])}_{\gamma_t^{\$}} \$_t^d + \underbrace{\delta E_t[\gamma_{t+1}^{a}]}_{\gamma_t^{a}} a_t^d + \underbrace{\left(\delta E_t[\gamma_{t+1}^{aa}] - f_t\left(1 + \delta R E_t[\gamma_{t+1}^{\$}]\right)\right)}_{\gamma_t^{aa}} (a_t^d)^2 + \underbrace{\delta E_t[\gamma_{t+1}]}_{\gamma_t} \tag{20}$$

Appendix A.2 Value function coefficient recursions

Appendix A.2.1 Maturity / reissuance

$$\gamma_T^{\$} = 1 - R^{-1} + \delta R E_T\left[\gamma_{T+1}^{\$}\right] \tag{21}$$
$$\gamma_T^{a} = \left(1 + \delta R E_T\left[\gamma_{T+1}^{\$}\right]\right) V_T^{\text{issuer}} \tag{22}$$
$$\gamma_T^{aa} = -\left(1 + \delta R E_T\left[\gamma_{T+1}^{\$}\right]\right) f_T \tag{23}$$
$$\gamma_T = \delta\left(E_T\left[\gamma_{T+1}\right] - \left(q_T^{\text{issuer}}\right)^2 E_T\left[\gamma_{T+1}^{aa}\right]\right) \tag{24}$$

Appendix A.2.2 Inter-dealer trade

$$\gamma_t^{\$,i} = 1 - R^{-1} + \delta R E_t\left[\gamma_{t+1}^{\$}\right] \tag{25}$$

$$\gamma_t^{a,i} = \left(1 + \delta R E_t\left[\gamma_{t+1}^{\$}\right]\right) p_t^i \tag{26}$$

$$\gamma_t^{aa,i} = -\left(1 + \delta R E_t\left[\gamma_{t+1}^{\$}\right]\right) f_t \tag{27}$$

$$\gamma_t^{i} = \delta\left(E_t\left[\gamma_{t+1}\right] - \left(a_{t+1}^{d}\right)^2 E_t\left[\gamma_{t+1}^{aa}\right]\right) \tag{28}$$

Appendix A.2.3 Customer-dealer trade

$$\gamma_t^{\$,c_j} = 1 - R^{-1} + \delta R E_t\left[\gamma_{t+1}^{\$}\right] \tag{29}$$

$$\gamma_t^{a,c_j} = \eta_j \delta E_t\left[\gamma_{t+1}^{a}\right] + \left(1 + \delta R E_t\left[\gamma_{t+1}^{\$}\right]\right)(1-\eta_j)\tilde{V}_t^{c_j} \tag{30}$$

$$\gamma_t^{aa,c_j} = \delta\eta_j E_t[\gamma_{t+1}^{aa}] - \left(1 + \delta R E_t\left[\gamma_{t+1}^{\$}\right]\right) f_t \tag{31}$$

$$\gamma_t^{c_j} = E_t\left[\delta\gamma_{t+1} - a_{t+1}^{d,c_j}\left(\delta\eta_j\left(\gamma_{t+1}^{a} + \gamma_{t+1}^{aa} a_{t+1}^{d,c_j}\right) + (1-\eta_j)\left(1 + \delta R\gamma_{t+1}^{\$}\right)\tilde{V}_t^{c_j}\right)\right] \tag{32}$$

Appendix A.2.4 No trading

$$\gamma_t^{\$} = 1 - R^{-1} + \delta R E_t[\gamma_{t+1}^{\$}] \tag{33}$$

$$\gamma_t^{a} = \delta E_t[\gamma_{t+1}^{a}] \tag{34}$$

$$\gamma_t^{aa} = \delta E_t[\gamma_{t+1}^{aa}] - f_t\left(1 + \delta R E_t[\gamma_{t+1}^{\$}]\right) \tag{35}$$

$$\gamma_t = \delta E_t[\gamma_{t+1}] \tag{36}$$

Appendix A.2.5 Uncertainty about trade type

$$V_t^d(\$_t^d, a_t^d, s_t) = \pi_i V_t^{d,i} + \sum_j \pi_j^c V_t^{d,c_j} + (1 - \pi_i - \pi_c) V_t^d \tag{37}$$

$$= \underbrace{\left(1 - R^{-1} + \delta R E_t\left[\gamma_{t+1}^{\$}\right]\right)}_{=\gamma_t^{\$}} \$_t^d + \underbrace{\left(\left(1 + \delta R E_t\left[\gamma_{t+1}^{\$}\right]\right)\left(\pi_i p_t^i + \sum_j \pi_j^c (1 - \eta_j) V_t^{c_j}\right) + \left(1 - \pi_i - \pi_c + \sum_j \pi_j^c \eta_j\right) \delta E_t\left[\gamma_{t+1}^a\right]\right)}_{=\gamma_t^a} a_t^d$$

$$+ \underbrace{\left(-\left(1 + \delta R E_t\left[\gamma_{t+1}^{\$}\right]\right) f_t + \delta\left(1 - \pi_i - \pi_c + \sum_j \pi_j \eta_j\right) E_t[\gamma_{t+1}^{aa}]\right)}_{=\gamma_t^{aa}} (a_t^d)^2 + \underbrace{\pi_i \gamma_t^i + \sum_j \pi_j^c \gamma_t^{c_j} + (1 - \pi_i - \pi_c)\gamma_t}_{=\gamma_t} \tag{38}$$

Appendix A.2.6 Value function coefficient recursions

$$\gamma_t^{\$} = 1 - R^{-1} + \delta R E_t\left[\gamma_{t+1}^{\$}\right] \tag{39}$$

$$\Rightarrow \gamma^{\$} = \frac{1 - R^{-1}}{1 - \delta R} \tag{40}$$

if $\delta R < 1$. So $\gamma^{\$}$ is a constant.

$$\begin{aligned}
\gamma_t^a &= \left(1+\delta R\gamma^{\$}\right)\left(\pi_i p_t^i + \sum_j \pi_j^c(1-\eta_j)V_t^{c_j}\right) + \delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right) E_t\left[\gamma_{t+1}^a\right] && (41)\\
&= \ ... \\
&= \left(1+\delta R\gamma^{\$}\right)\left(\sum_{u=0}^{T-t-1}\left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right)\right)^u\left(\pi_i E_t[p_{t+u}^i] + \sum_j \pi_j^c(1-\eta_j)E_t[\tilde{V}_{t+u}^{c_j}]\right)\right. \\
&\qquad \left. + \left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right)\right)^{T-t} E_t\left[V_T^{\text{issuer}}\right]\right) && (42)
\end{aligned}$$

which only depends on the state at time t.

$$\begin{aligned}
\gamma_t^{aa} &= -\left(1+\delta R\gamma^{\$}\right) f_t + \delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right) E_t[\gamma_{t+1}^{aa}] && (43)\\
&= \ ... \\
&= -\left(1+\delta R\gamma^{\$}\right)\sum_{u=0}^{T-t}\left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right)\right)^u E_t[f_{t+u}] && (44)
\end{aligned}$$

which only depends on the state at time t.

Appendix A.3 Prices and quantities

Appendix A.3.1 Customer-dealer trade

The dealer's post-trade inventory position, after buying q_t^c units of the risky asset from a customer, is given by:

$$
\begin{aligned}
a_{t+1}^{d,c} &\equiv a_t^d + q_t^c && (45)\\
&= \frac{\left(1+\delta R E_t[\gamma_{t+1}^{\$}]\right)\tilde{V}_t^c - \delta E_t\left[\gamma_{t+1}^a\right]}{2\delta E_t\left[\gamma_{t+1}^{aa}\right]} && (46)\\
&= \frac{\delta E_t\left[\sum_{u=0}^{T-t-2}(\delta(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j))^u(\pi_i p_{t+u+1}^i + \sum_j \pi_j^c(1-\eta_j)\tilde{V}_{t+u+1}^{c_j}) + (\delta(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j))^{T-t-1}V_T^{\text{issuer}}\right] - V_t^c}{2\delta E_t\left[\sum_{u=0}^{T-t-1}\left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right)\right)^u f_{t+u+1}\right]} && (47)
\end{aligned}
$$

The dealer's new position after trading with the customer does not depend on her inventory before.

Appendix A.3.2 Economic decomposition of customer-dealer price

I showed earlier that customer-dealer price is a weighted average of the customer and dealer valuations, where the weights are given by their bargaining powers. The dealer's valuation is the price such that their gain from trade is zero. I now find a recursion for the dealer's gain from trade which gives intuition for the customer-dealer price. The gain from trade for a dealer with a_t^d units of inventory of the risky asset from buying q_t^c units at price per

unit p_t^c is:

$$
\begin{aligned}
g_t^d(a_t^d, p_t^c, q_t^c) &= q_t^c E_t\left[\delta\gamma_{t+1}^a + \delta\left(2a_t^d + q_t^c\right)\gamma_{t+1}^{aa} - (1+\delta R\gamma^{\$})p_t^c\right] && (48)\\
&= (1+\delta R\gamma^{\$})\left(-p_t^c q_t^c + \frac{\delta E_t\left[\gamma_{t+1}^a + \left(2a_t^d + q_t^c\right)\gamma_{t+1}^{aa}\right] q_t^c}{1+\delta R\gamma^{\$}}\right) && (49)\\
&= (1+\delta R\gamma^{\$})\left(-p_t^c q_t^c - \delta\left(2a_t^d + q_t^c\right) q_t^c E_t[f_{t+1}] + \delta\pi_i E_t[p_{t+1}^i] q_t^c\right. && (50)\\
&\quad \left. +\delta\left((1-\pi_i-\pi_c)E_t\underbrace{\left[\frac{\delta E_{t+1}\left[\gamma_{t+2}^a + \left(2a_t^d + q_t^c\right)\gamma_{t+2}^{aa}\right] q_t^c}{1+\delta R\gamma^{\$}}\right]}_{=\frac{g_{t+1}^d(a_t^d, p_t^c, q_t^c)}{1+\delta R\gamma^{\$}} + p_t^c q_t^c} + \sum_j \pi_j^c E_t\underbrace{\left[\eta_j\left(\frac{\delta E_{t+1}\left[\gamma_{t+2}^a + \left(2a_t^d + q_t^c\right)\gamma_{t+2}^{aa}\right]}{1+\delta R\gamma^{\$}}\right) + (1-\eta_j)V_{t+1}^{c_j}\right]}_{=p_{t+1}^{c_j}(a_t^d, q_t^c)} q_t^c\right)\right)
\end{aligned}
$$

using the recursions for the γ_t^a and γ_t^{aa}. We get the recursion:

$$
\begin{aligned}
g_t^d\left(a_t^d, p_t^c, q_t^c\right) &= (1+\delta R\gamma^{\$})\left(-p_t^c q_t^c - \delta\left(2a_t^d + q_t^c\right) q_t^c E_t[f_{t+1}] + \delta\pi_i E_t[p_{t+1}^i] q_t^c + \delta\left((1-\pi_i-\pi_c)p_t^c q_t^c + \sum_j \pi_j^c E_t\left[p_{t+1}^{c_j}(a_t^d, q_t^c)\right] q_t^c\right)\right)\\
&\quad +\delta(1-\pi_i-\pi_c)E_t[g_{t+1}^d(a_t^d, p_t^c, q_t^c)] && (51)
\end{aligned}
$$

Iterating:

$$
\begin{aligned}
g_t^d\left(a_t^d, p_t^c, q_t^c\right) &= (1+\delta R\gamma^{\$}) q_t^c \sum_{u=0}^{T-t-2}(\delta(1-\pi_i-\pi_c))^u\left(-p_t^c - \delta\left(2a_t^d + q_t^c\right)E_t[f_{t+u+1}] + \delta\pi_i E_t[p_{t+u+1}^i] + \delta((1-\pi_i-\pi^c)p_t^c\right. && (52)\\
&\quad \left. + \sum_j \pi_j^c E_t\left[p_{t+u+1}^{c_j}(a_t^d, q_t^c)\right]\right) + (\delta(1-\pi_i-\pi_c))^{T-t-1} E_t[g_{T-1}^d(a_t^d, q_t^c, p_t^c)]
\end{aligned}
$$

Using the boundary condition from the last trade before the bond matures:

$$g^d_{T-1}(a^d_t, q^c_t, p^c_t) = (1+\delta R\gamma^{\$})\left(-p^c_t q^c_t + \delta E_{T-1}[V^{\text{issuer}}_T]q^c_t - \delta E_{T-1}[f_T](2a^d_t + q^c_t)q^c_t\right) \tag{53}$$

we get

$$\begin{aligned} g^d_t\left(a^d_t, p^c_t, q^c_t\right) = (1+\delta R\gamma^{\$})q^c_t\Bigg(&-p^c_t - \delta\left(2a^d_t + q^c_t\right)\sum_{u=0}^{T-t-1}\left(\delta(1-\pi_i-\pi_c)\right)^u E_t[f_{t+u+1}] \\ &+\delta\sum_{u=0}^{T-t-2}\left(\delta(1-\pi_i-\pi_c)\right)^u\left(\pi_i E_t[p^i_{t+u+1}] + \sum_j \pi^c_j E_t\left[p^{c_j}_{t+u+1}(a^d_t, q^c_t)\right]\right) + \left(\delta(1-\pi_i-\pi_c)\right)^{T-t-1}\delta E_t\left[V^{\text{issuer}}_T\right]\Bigg) \end{aligned} \tag{54}$$

From earlier, we know we can write $p^c_t\left(a^d_t, q^c_t\right) = \left(2a^d_t + q^c_t\right)x_t + y_t$, where x_t and y_t do not depend on a^d_t or q^c_t. Let q^c_t denote the optimal quantity the dealer would buy if she met a customer at time t. Let q^c_{t+u+1} be the optimal quantity the dealer would buy if she met a customer at time $t+u+1$, assuming her last trade was buying q_t at time t. Then we can write:

$$q^c_t \times p^{c_j}_{t+u+1}\left(a^d_t, q^c_t\right) = \underbrace{-q^c_{t+u+1} \times p^{c_j}_{t+u+1}\left(a^d_t + q^c_t, q^c_{t+u+1}\right)}_{\text{trade at } t \text{ and } t+u+1} - \underbrace{\left(-\left(q^c_t + q^c_{t+u+1}\right) \times p^{c_j}_{t+u+1}\left(a^d_t, q^c_t + q^c_{t+u+1}\right)\right)}_{\text{trade at } t+u+1 \text{ only}} \tag{55}$$

which is the cash gain from trade at time $t+u+1$, if dealer had traded at time t. Let q^i_{t+u+1}, and q^{issuer}_T denote the optimal quantity the dealer would buy if she meets a dealer or customer at time $t+u+1$, or issuer at time T respectively, assuming her last trade was buying q^c_t at time t. If a dealer

has inventory a_t^d, then her gain from trade, $g_t^d\left(a_t^d, p_t^c, q_t^c\right)$, from buying q_t^c units at price per unit $p_t^c(a_t^d, q_t^c)$, is:

$$
\begin{aligned}
&\underbrace{-p_t^c q_t^c}_{\text{buy } q_t^c \text{ units from time-}t \text{ customer}} \\
&\underbrace{-\delta \overbrace{\left(2a_t^d + q_t^c\right)}^{(a_t^d+q_t^c)^2-(a^d)^2} q_t^c \sum_{u=0}^{T-t-1} \left(\delta(1-\pi_i-\pi_c)\right)^u E_t[f_{t+u+1}]}_{\text{discounted costs on } (a_t^d+q_t^c) \text{ units instead of } a_t^d, \text{ until next expected trade}} \\
&+\delta \sum_{u=0}^{T-t-2} \left(\delta(1-\pi_i-\pi_c)\right)^u \underbrace{\left(\pi_i E_t[p_{t+u+1}^i]\left(\cancel{-q_{t+u+1}^i} - \left(-(\cancel{q_{t+u+1}^i} + q_t^c)\right)\right) + \sum_j \pi_j^c E_t \left[\underbrace{-q_{t+u+1}^{c_j} . p_{t+u+1}^{c_j}\left(a_t^d + q_t^c, q_{t+u+1}^{c_j}\right)}_{\text{trade at } t \text{ and } t+u+1} - \underbrace{\left(-\left(q_t^c + q_{t+u+1}^{c_j}\right) . p_{t+u+1}^c\left(a_t^d, q_t^c + q_{t+u+1}^{c_j}\right)\right)}_{\text{trade at } t+u+1 \text{ only}} \right] \right)}_{\text{buy } q_{t+u+1} \text{ units from time-}(t+u+1) \text{ counterparty, instead of } q_{t+u+1}+q_t^c. \text{ The counterparty is a dealer with probability } \pi_i, \text{ and customer } j \text{ with probability } \pi_j^c.} \\
&+ \underbrace{\left(\delta(1-\pi_i-\pi_c)\right)^{T-t-1} \delta \left(\cancel{a_t^d} + q_t^c - (\cancel{(a_t^d)})\right) E_t\left[V_T^{\text{issuer}}\right]}_{\substack{\text{cash in } a_t^d+q_t^c \text{ units instead of } a_t^d \text{ from issuer at time } T, \text{ if dealer doesn't} \\ \text{meet a counterparty in the previous } T-t-1 \text{ rounds of trading}}}
\end{aligned}
\tag{56}
$$

all multiplied by $(1+\delta R\gamma^{\$})$. I showed earlier that the customer-dealer price can be expressed as a weighted average of the dealer's and customer's valuations, where the weights are given by their relative bargaining power. The dealer's valuation can be characterized as the customer-dealer price

such that the dealer's gain from trade at time t is zero. The customer-dealer price per unit can thus be written recursively:

$$
\begin{aligned}
p_t^c \;=\; \eta \Bigg(& \underbrace{-\delta\left(2a_t^d + q_t^c\right) \sum_{u=0}^{T-t-1} \left(\delta(1-\pi_i-\pi_c)\right)^u E_t[f_{t+u+1}]}_{\text{inventory costs on } (a^d+q^c) \text{ units instead of } a^d\text{, until next trade}} \\
& + \delta \sum_{u=0}^{T-t-2} \left(\delta(1-\pi_i-\pi_c)\right)^u \underbrace{\Bigg(\pi_i E_t[p_{t+u+1}^i] + \sum_j \pi_j^c E_t \Bigg[\underbrace{-q_{t+u+1}^{c_j} . p_{t+u+1}^{c_j}\left(a_t^d + q_t^c, q_{t+u+1}^{c_j}\right)}_{\text{trade at } t \text{ and } t+u+1} - \underbrace{\left(-\left(q_t^c + q_{t+u+1}^{c_j}\right) . p_{t+u+1}^c\left(a_t^d, q_t^c + q_{t+u+1}^{c_j}\right)\right)}_{\text{trade at } t+u+1 \text{ only}} \Bigg] / q_t^c \Bigg)}_{\substack{\text{buy } q_{t+u+1} \text{ units from time-}(t+u+1) \text{ counterparty, instead of } q_{t+u+1}+q_t^c\text{. The counterparty is a dealer with} \\ \text{probability } \pi_i\text{, and customer } j \text{ with probability } \pi_j^c}} \\
& + \underbrace{\left(\delta(1-\pi_i-\pi_c)\right)^{T-t-1} \delta E_t\left[V_T^{\text{issuer}}\right]}_{\substack{\text{cash in } a_t^d+q_t^c \text{ units instead of } a_t^d \text{ from issuer at time } T\text{, if} \\ \text{dealer doesn't meet a counterparty in the previous } T-t-1 \\ \text{rounds of trading}}} \Bigg) \\
& + (1-\eta)\tilde{V}_t^c
\end{aligned}
\tag{57}
$$

Appendix A.3.3 Inter-dealer trade

Using equations (6), (42), and (44), the dealer's post-trade inventory position, after buying q_t^i units of the risky asset from the inter-dealer market, is given by:

$$a_{t+1}^{d,i} = \frac{\delta E_t^d \left[\sum_{u=0}^{T-t-2}(\delta(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j))^u(\pi_i p_{t+u+1}^i+\sum_j \pi_j^c(1-\eta_j)V_{t+u+1}^{c_j})+(\delta(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j))^{T-t-1}V_T^{\text{issuer}}\right]-p_t^i}{2\delta E_t^d\left[\sum_{u=0}^{T-t-1}(\delta(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j))^u f_{t+u+1}\right]} \tag{58}$$

Market-clearing in the inter-dealer market states that net trading must be zero. Let a_t^D denote the aggregate inventory of dealers who trade in the inter-dealer market at time t. Each dealer will hold the same inventory after trading, as they have identical preferences. $\pi_i/\tilde{\pi}_i$ is the proportion of dealers trading in the inter-dealer market.

$$a_{t+1}^{d,i} = \frac{a_t^D}{\pi_i/\tilde{\pi}_i} \tag{59}$$

This gives a recursion for the inter-dealer price, which can be solved by backwards induction from the trade before the maturity of the asset.

$$\begin{aligned} p_t^i = {} & \delta E_t\left[\sum_{u=0}^{T-t-2}\left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right)\right)^u\left(\pi_i p_{t+u+1}^i+\sum_j \pi_j^c(1-\eta_j)V_{t+u+1}^{c_j}\right)+\left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right)\right)^{T-t-1}V_T^{\text{issuer}}\right] \\ & -\frac{2\delta a_t^D}{\pi_i/\tilde{\pi}_i}E_t\left[\sum_{u=0}^{T-t-1}\left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right)\right)^u f_{t+u+1}\right] \end{aligned} \tag{60}$$

Appendix A.3.4 Economic decomposition of inter-dealer price

Using equations (6), (41), and (43), the inter-dealer price can be written as:

$$p_t^i = -2\delta\bar{a}_t^{d,i}E_t[f_{t+1}] + \delta\pi_i E_t[p_{t+1}^i] + \delta(1-\pi_i-\pi_c)\overbrace{\frac{2\delta\bar{a}_t^{d,i}E_t[\gamma_{t+2}^{aa}] + \delta E_t[\gamma_{t+2}^{a}]}{1+\delta R\gamma^{\$}}}^{E_t[p_{t+1}^i]} + \delta\sum_j \pi_j^c \left(\overbrace{\frac{\eta_j\delta E_t[2\bar{a}_t^{d,i}\gamma_{t+2}^{aa}] + \gamma_{t+2}^{a}}{1+\delta R\gamma^{\$}} + (1-\eta_j)V_{t+1}^{c_j}}^{p_{t+1}^{c_j}(a_t^{d,i},q_t^i) + \frac{\eta_j\delta q_t^i E_t[\gamma_{t+2}^{aa}]}{1+\delta R\gamma^{\$}}} \right) \tag{61}$$

$$= \delta\sum_{u=0}^{T-t-2}(\delta(1-\pi_i-\pi_c))^u E_t\left[-2\bar{a}_t^{d,i}f_{t+u+1} + \pi_i p_{t+u+1}^i + \sum_j \pi_j^c \left(p_{t+u+1}^{c_j}(a_t^{d,i},q_t^i) + \frac{\eta_j\delta q_{t+u+1}^i\gamma_{t+2}^{aa}}{1+\delta R\gamma^{\$}} \right) \right.$$
$$\left. +(\delta(1-\pi_i-\pi_c))^{T-t-1}p_{T-1}^i \right] \tag{62}$$

by iterating forward. Making the following substitutions:

- $p_{t+u+1}^{c_j}(a_t^d,q_t^i) + \frac{\eta_j\delta q_t^i E_t[\gamma_{t+1}^{aa}]}{1+\delta R\gamma^{\$}} = \frac{\partial}{\partial q_t^i}\left[q_t^i p_{t+u+1}^{c_j}(a_t^d,q_t^i)\right]$
- $q_t^i p_{t+u+1}^{c_j}\left(a_t^d,q_t^i\right) = \underbrace{-q_{t+u+1}^i \times p_{t+u+1}^{c_j}\left(a_t^d+q_t^i,q_{t+u+1}^i\right)}_{\text{trade at } t \text{ and } t+u+1} \underbrace{-\left(-\left(q_t^i+q_{t+u+1}^i\right)\times p_{t+u+1}^{c_j}\left(a_t^d,q_t^i+q_{t+u+1}^i\right)\right)}_{\text{trade at } t+u+1 \text{ only}}$
- $p_{T-1}^i = \frac{\delta E_{T-1}[V_T^{\text{issuer}}] - 2\bar{a}_T^{d,i}E_{T-1}[f_T]}{1+\delta R\gamma^{\$}}$

this can be written further as:

$$
\begin{aligned}
p_t^i(\bar{a}_t^{d,i}) \quad = \quad & \frac{\partial}{\partial q_t^i}\Bigg[\underbrace{-\delta(\bar{a}_t^{d,i})^2\sum_{u=0}^{T-t-1}(\delta(1-\pi_i-\pi_c))^u E_t[f_{t+u+1}]}_{\text{discounted inventory costs until next expected trade}}+ \\
& \underbrace{\sum_{u=0}^{T-t-2}(\delta(1-\pi_i-\pi_c))^u(\pi_i E_t[p^i_{t+u+1}]((q^i_{t+u+1}+q^i_t)-q^i_{t+u+1})+\sum_j \pi^c_j((q^i_t+q^c_{t+u+1}).p^{c_j}_{t+u+1}(a^{d,i}_t,q^i_t+q^c_{t+u+1})-q^c_{t+u+1}.p^{c_j}_{t+u+1}(a^{d,i}_t+q^i_t,q^c_{t+u+1})))}_{\text{sell } q_{t+u+1} \text{ units instead of } q_{t+u+1}+q^i_t \text{ at next trade. Next counterparty is dealer with probability } \pi_i \text{ and customer } j \text{ with probability } \pi^{c_j}.} \\
& + \quad \underbrace{(\delta(1-\pi_i-\pi_c))^{T-t-1}E_t\left[V_T^{\text{issuer}}\right]\left((a^d_t+q^i_t)-a^d_t\right)]}_{\substack{\text{cash in } a^d_t \text{ units instead of } a^d_t+q^i_t \text{ from issuer at time } T, \text{ if dealer doesn't} \\ \text{meet a counterparty in the previous } T-t-1 \text{ rounds of trading}}}
\end{aligned} \tag{63}
$$

Appendix A.3.5 Issuance

The dealer's inventory position after maturity of one risky asset, and issuance of a new one, is given by:

$$a^d_{T_1+1} = q^{\text{issuer}}_{T_1} =$$

$$\frac{p^{\text{issuer}}_{T_1}\left(1+\delta R E_T[\gamma^{\$}_{T+1}]\right)-\delta E_T\left[\gamma^a_{T+1}\right]}{2\delta E_T\left[\gamma^{aa}_{T+1}\right]} = \tag{64}$$

$$\frac{\delta E_{T_1}\left[\sum_{u=0}^{T_2-T_1-2}(\delta(1-\pi_i-\pi_c+\sum_j \pi^c_j\eta_j))^u(\pi_i p^i_{T_1+u+1}+\sum_j \pi^c_j(1-\eta_j)\tilde{V}^{c_j}_{T_1+u+1})+(\delta(1-\pi_i-\pi_c+\sum_j \pi^c_j\eta_j))^{T_2-T_1-1}V^{\text{issuer}}_{T_2}\right]-p^{\text{issuer}}_{T_1}}{2\delta E_{T_1}\left[\sum_{u=0}^{T_2-T_1-1}\left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi^c_j\eta_j\right)\right)^u f_{T_1+u+1}\right]} \tag{65}$$

Appendix A.4 Markup

From equation (17) the customer-dealer price is given by:

$$p_t^c = \eta\left(\frac{\delta E_t[\gamma_{t+1}^a] + \delta(2a_t^{d,c} + q_t^c)E_t[\gamma_{t+1}^{aa}]}{1+\delta R\gamma^{\$}}\right) + (1-\eta)\tilde{V}_t^c \tag{66}$$

$$= \tilde{V}_t^c - \frac{\eta\delta q_t^c E_t[\gamma_{t+1}^{aa}]}{1+\delta R\gamma^{\$}} \tag{67}$$

using equation (18), which becomes:

$$p_t^c = \tilde{V}_t^c - \eta\delta q_t^c \sum_{u=0}^{T-t}\left(\delta\left(1-\pi_i-\pi_c+\sum_j \pi_j^c\eta_j\right)\right)^u E_t[f_{t+u}] \tag{68}$$

using equation (44).

From equation (6) the inter-dealer price is:

$$p_t^i = \frac{\delta E_t^d\left[\gamma_{t+1}^a\right] + 2\delta(a_t^{d,i} + q_t^i)E_t^d\left[\gamma_{t+1}^{aa}\right]}{1+\delta R\gamma^{\$}} \tag{69}$$

$$= \tilde{V}_t^c + \frac{2\delta((a_t^{d,i} + q_t^i) - (a_t^{d,c} + q_t^c))E_t^d\left[\gamma_{t+1}^{aa}\right]}{1+\delta R\gamma^{\$}} \tag{70}$$

using equation (18).

The markup when the dealer is selling to the customer is thus given by:

$$p_t^c - p_t^i = \left(\cancel{\tilde{Y}_t^c} - \frac{\eta\delta q_t^c E_t[\gamma_{t+1}^{aa}]}{1+\delta R\gamma^{\$}}\right) - \left(\cancel{\tilde{Y}_t^c} + \frac{2\delta((a_t^{d,i}+q_t^i)-(a_t^{d,c}+q_t^c))E_t^d\left[\gamma_{t+1}^{aa}\right]}{1+\delta R\gamma^{\$}}\right) \tag{71}$$

$$= \delta\left(\eta q_t^c - 2\left(\underbrace{(a_t^{d,c}+q_t^c)}_{\substack{\text{dealer's post-trade}\\ \text{inventory}}} - \underbrace{(a_t^{d,i}+q_t^i)}_{\substack{\text{dealers' mean inventory}\\ \text{in inter-dealer market}}}\right)\right)\sum_{u=0}^{T-t-1}\left(\delta\left(1-\pi_i-\sum_j \pi_j^c(1-\eta_j)\right)\right)^u E_t[f_{t+u+1}] \tag{72}$$

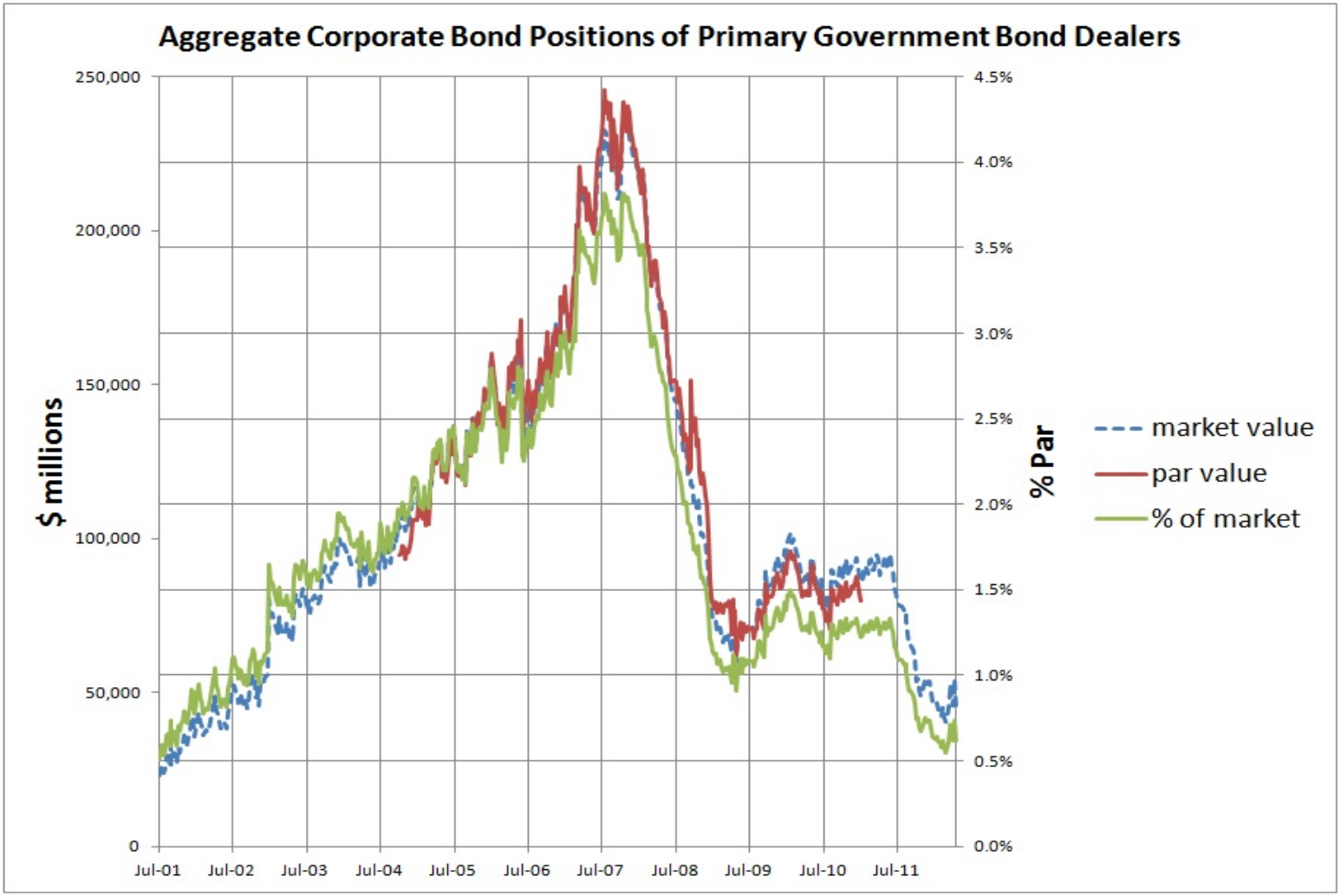

Figure 1: **Aggregate US corporate securities holdings by primary dealers.** This figure shows the time series of aggregate US corporate securities inventory held by primary government bond dealers, measured in market value, estimated par value, and an upper bound for % par of the corporate bond market. There is a large decrease in inventory starting from October 2007, which coincided with some dealing banks (Citi, Merrill Lynch, UBS) experiencing large asset write-downs. A second decrease in inventory began in June 2011, during the sovereign debt crisis. Data is from The Federal Reserve Bank of New York, which reports weekly holdings and trade volume, aggregated across all primary government bond dealers and across not just corporate bonds, but other corporate securities, which is why the par value is only estimated and the % par represents an upper bound. Table 1 of Randall (2015) lists the names of primary government bond dealers, and the major dealers who trade on MarketAxess, an electronic auction market for corporate bonds. There is a large overlap in the two lists, so it seems reasonable that the aggregate inventory of primary government bond dealers is likely representative of the wider dealer market.

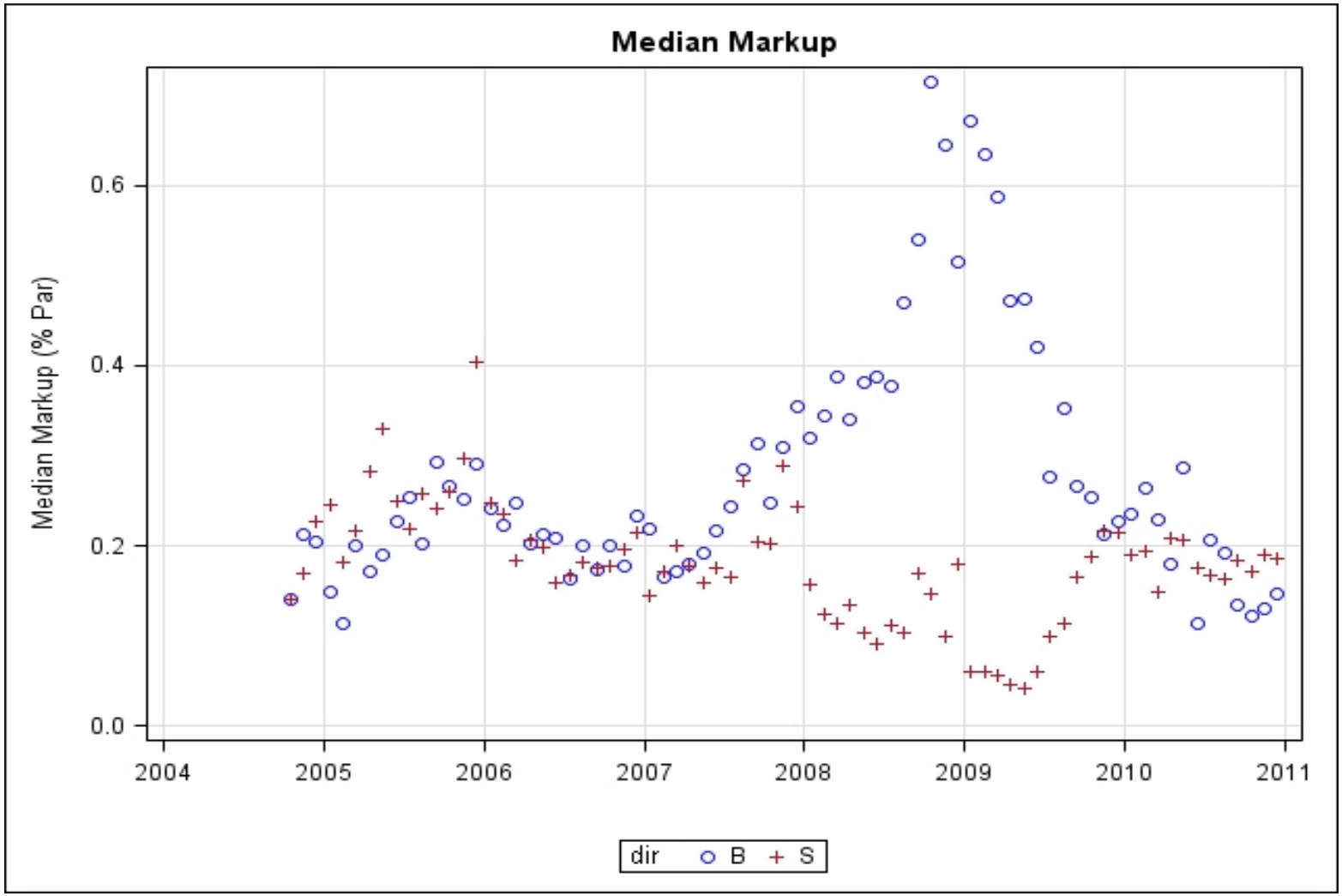

Figure 2: **Median markup over time.** This plot shows the time series of the median "markup", as a percentage of par, when the dealer is buying from a customer (B, blue circle) and when selling to a customer (S, red cross). The markup is defined as the difference between a customer-dealer (retail) price and the inter-dealer (wholesale) price. The bid-ask spread is the sum of the markup to buy and the markup to sell. In the financial crisis of 2007-09, as dealers sought to reduce their inventory, they priced bonds low to encourage customers to buy from them, and discourage them to sell to them. We see this effect by higher markups when dealers were buying, and lower when they were selling.

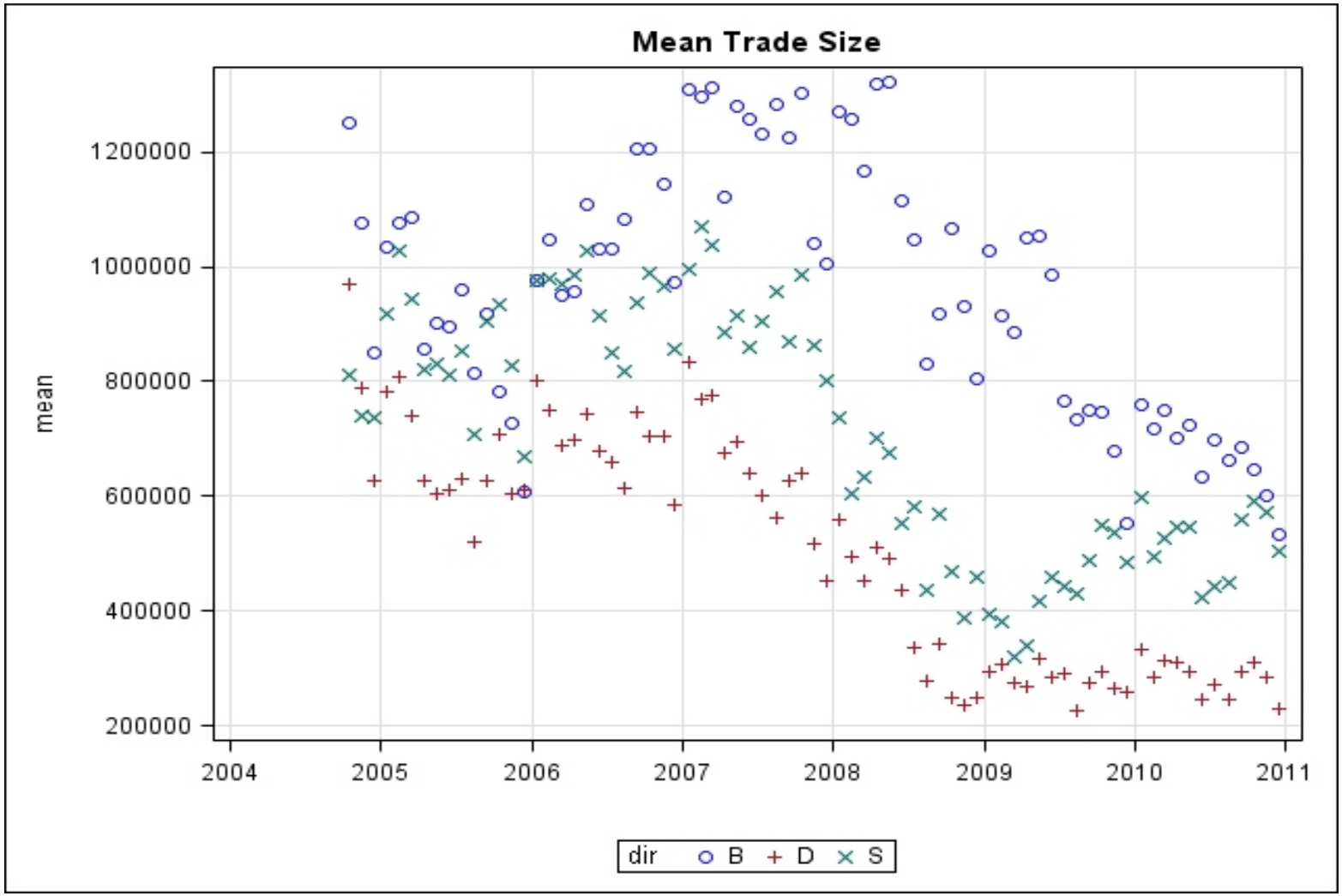

Figure 3: **Mean trade size over time.** This plot shows the time series of the mean trade size when the dealer is buying from a customer (B, blue circle), when selling to a customer (S, green diagonal cross), and for interdealer trades (D, red upright cross). During the financial crisis of 2007-09 average trade size decreased for all three categories of trade, and remained lower afterwards.

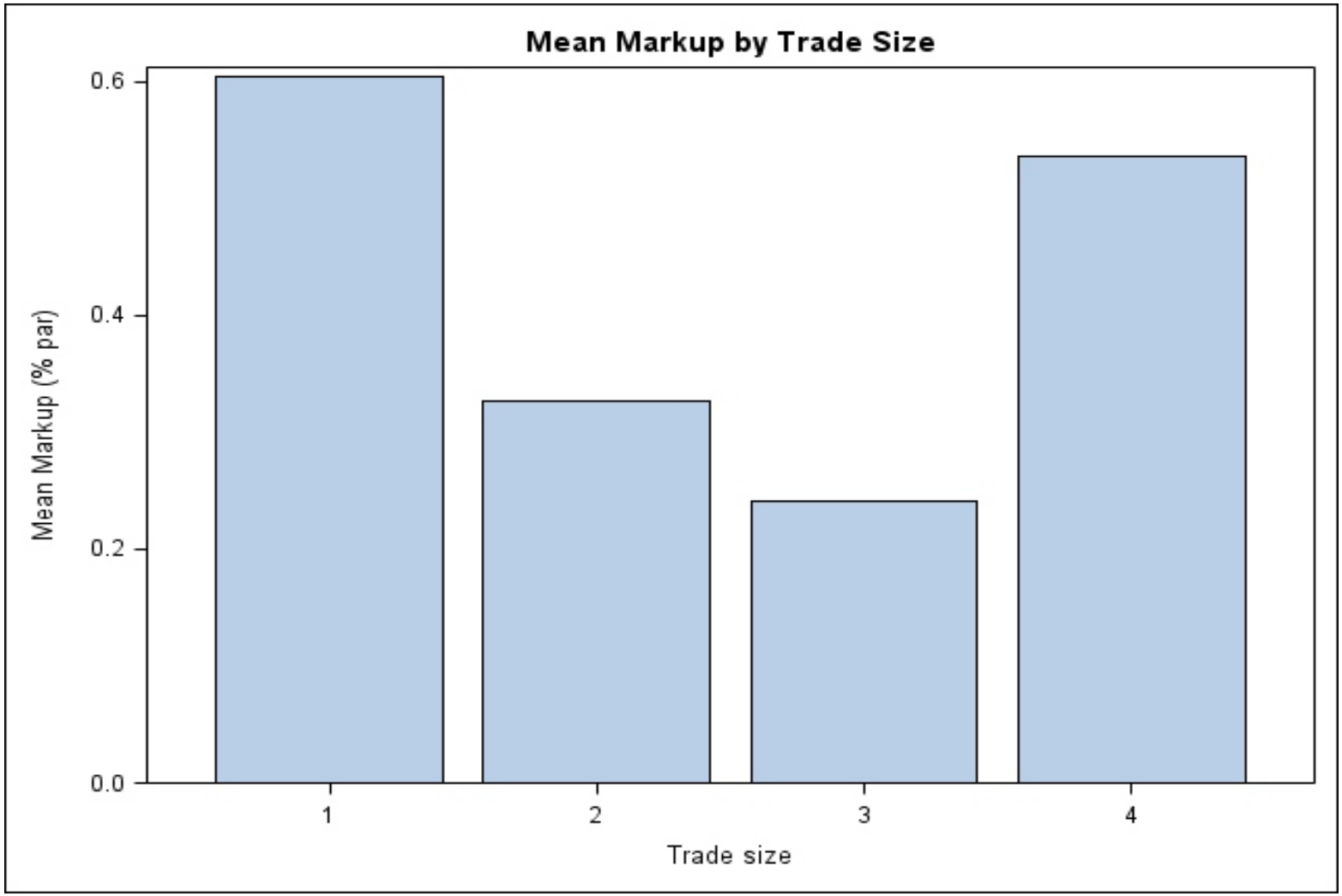

Figure 4: **Markup by trade size group.** This figure shows the mean markup, as a percentage of par, split into four trade size buckets: *retail* trades, which are classified in the previous empirical literature as having a par value of less than $100,000; *odd lot* trades with a par value of at least $100,000 but less than $1 million; *institutional* trades of at least $1 million, which is a round lot in the US corporate bond market, but less than $10 million; and trades of at least $10 million which I denote as *mega*.

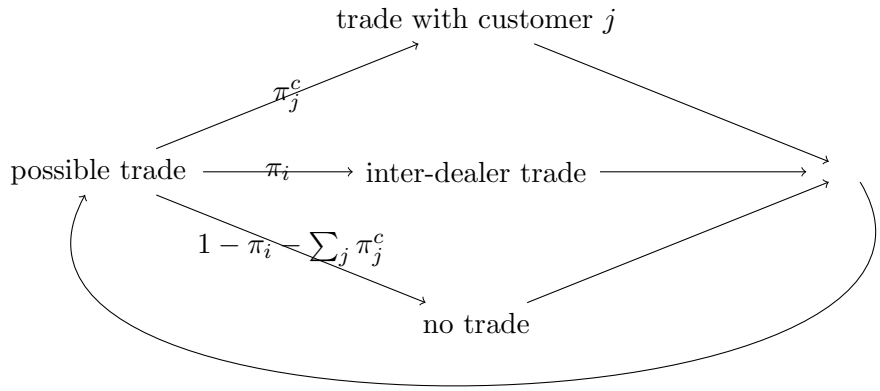

Figure 5: **Dealer time-line.** *π_j^c is the exogenous probability that each dealer trades with a customer of type j at time t. π_i is the exogenous probability that each dealer trades in the interdealer market at time t. $1 - \pi_i - \sum_j \pi_j^c$ is therefore the exogenous probability that each dealer does not trade that period. The probabilities of trade are independent across time. This cycle repeats until the risky asset matures, at which point asset holders receive a payoff, and a new risky asset is issued.*

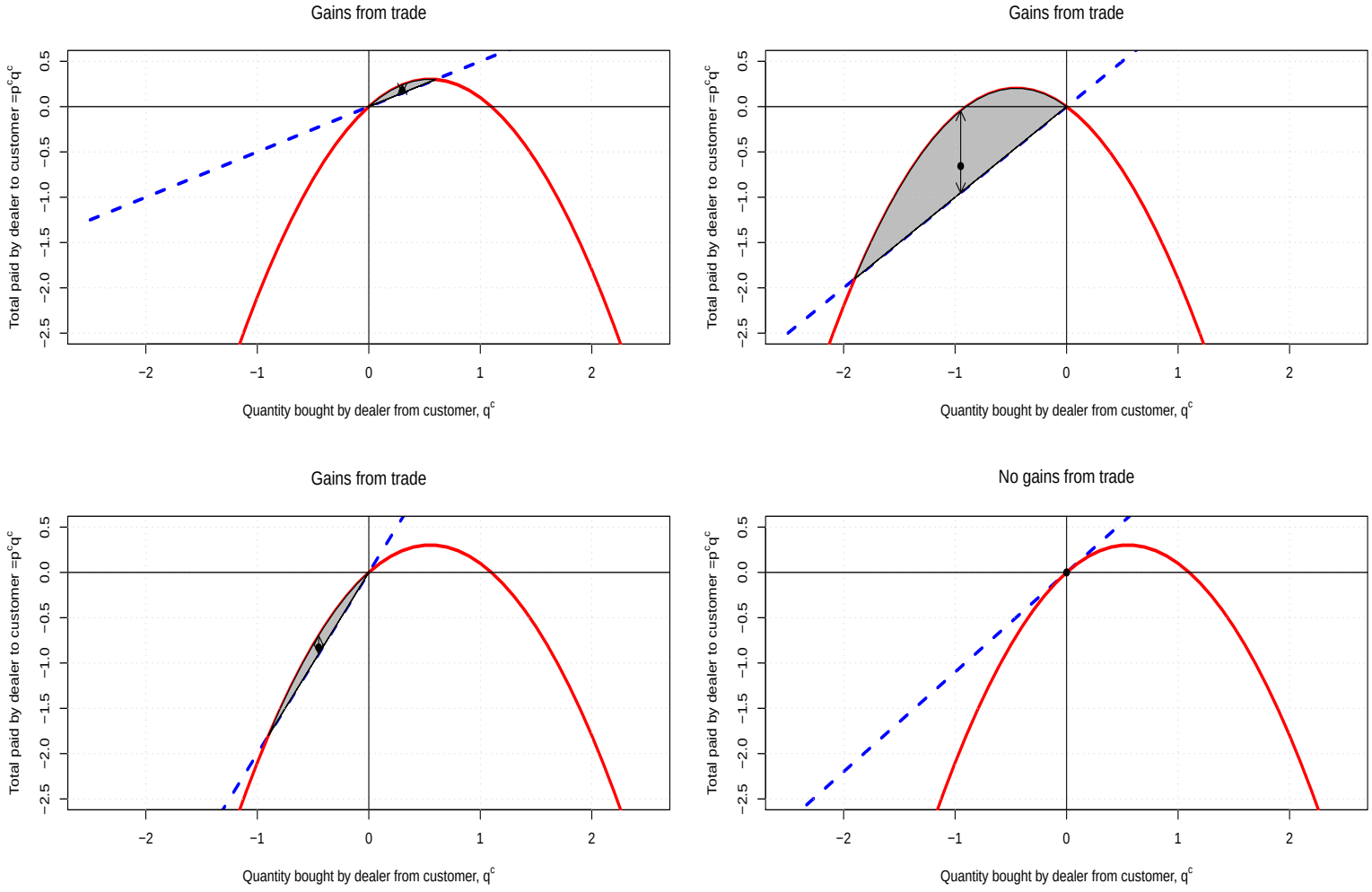

Figure 6: **Nash bargaining in dealer-customer trade.** The solid red and dashed blue lines represent the indifference curves for the dealer and customer, respectively. The region below (above) the solid (dashed) line represents the region where the dealer (customer) has a gain from trade. The intersection of these regions, where there are gains from trade to both dealer and customer, are shaded gray. The dot is plotted at the coordinates of the optimal quantity and total price (quantity $\times$ price/unit). The length of the arrows represent the gains from trade to the customer and dealer. Here the customer's relative bargaining power, η is $\frac{1}{3}$.

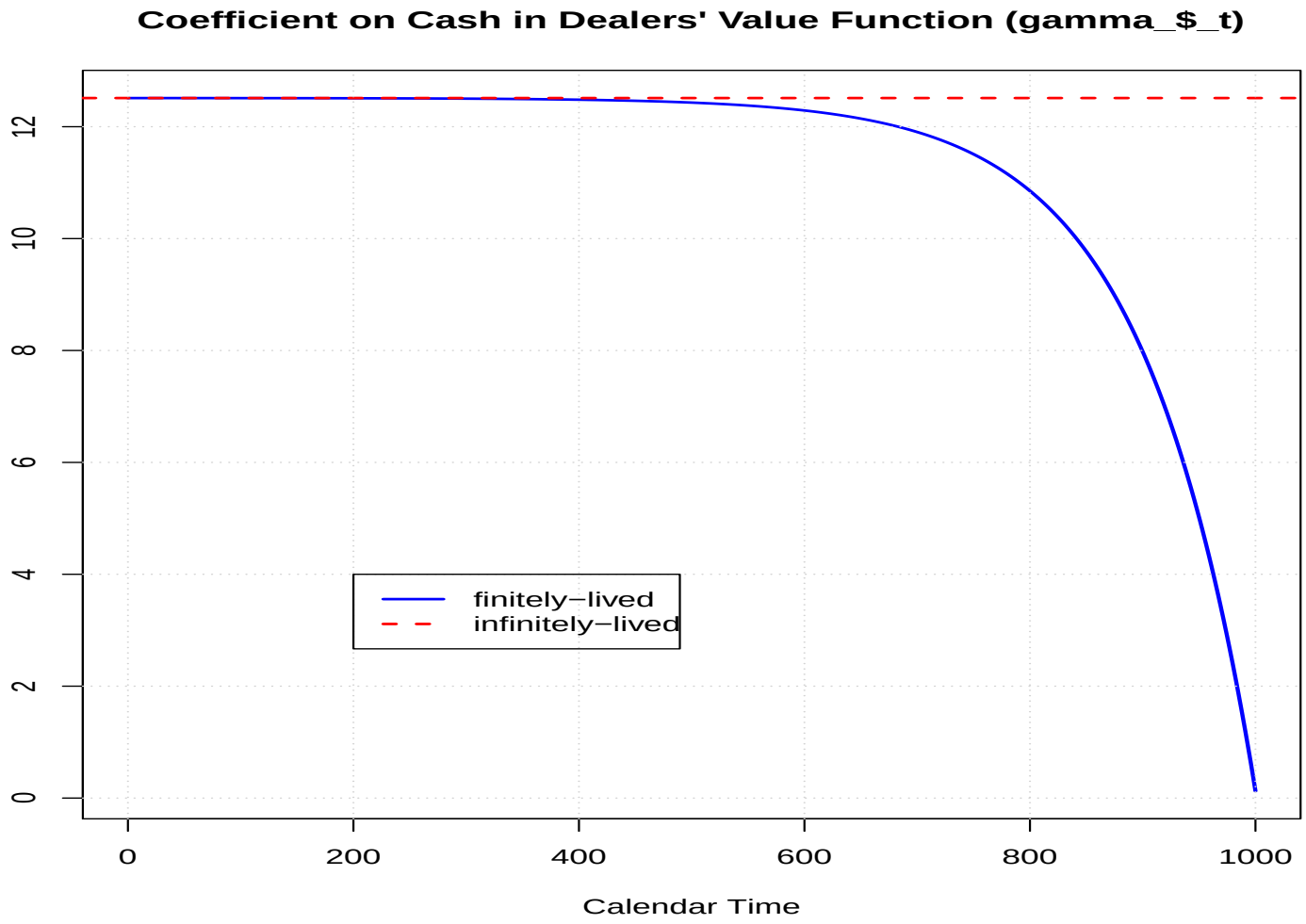

Figure 7: **Coefficient on Cash in Dealers' Value Function ($\gamma_t^{\$}$).** This figure shows the coefficient on the dealers' current cash position in their value function ($\gamma_t^{\$}$). The solid line is that coefficient when dealers are finitely-lived, dying at the maturity of the bond, and is deterministic but not constant over time. The dashed line is the coefficient when dealers are infinitely-lived, and is also the limit of the coefficient in the finitely-lived case as time to dealer death tends to infinity. In this example, there are 1,000 units of calendar time, representing dealers earn interest at 5% per year, and the discount factor, δ, is 0.99 per year.