FEDERAL RESERVE BANK
PHILADELPHIA

Working Papers
RESEARCH DEPARTMENT

Inventory, Market Making, and Liquidity in OTC Markets

Assa Cohen
Yeshiva University, Sy Syms School of Business

Mahyar Kargar
University of Illinois Urbana-Champaign

Benjamin Lester
Federal Reserve Bank of Philadelphia

Pierre-Olivier Weill
UCLA, NBER, and CEPR

WP 24-22

PUBLISHED
December 2024

ISSN: 1962-5361

DOI: https://doi.org/10.21799/frbp.wp.2024.22

Inventory, Market Making, and Liquidity in OTC Markets*

Assa Cohen† Mahyar Kargar‡ Benjamin Lester§ Pierre-Olivier Weill¶

October 8, 2024

Abstract

We develop a search-theoretic model of a dealer-intermediated over-the-counter market. Our key departure from the literature is to assume that, when a customer meets a dealer, the dealer can sell only assets that it already owns. Hence, in equilibrium, dealers choose to hold *inventory*. We derive the equilibrium relationship between dealers' costs of holding assets on their balance sheets, their optimal inventory holdings, and various measures of liquidity, including bid-ask spreads, trade size, volume, and turnover. Using transaction-level data from the corporate bond market, we calibrate the model to quantitatively assess the impact of post-crisis regulations on dealers' inventory costs, liquidity, and welfare.

Keywords: Over-the-counter markets, intermediation, liquidity, dealer inventory, financial regulation

JEL Classification: G11, G12, G21.

*We thank Jack Bao, Briana Chang, Tim Johnson, Charlie Kahn, Ricardo Lagos, Lucie Lebeau, Guillermo Ordoñez, Mariano Palleja, Dejanir Silva, James Vickery; two anonymous referees; conference participants at SED 2021, SFS Cavalcade 2022, AEA 2023, the 2023 Search and Matching Conference, the Search and Matching in Macro and Finance Virtual Seminar Series, the Virtual Finance Theory Seminar; and seminar participants at the Federal Reserve Bank of Atlanta, the Federal Reserve Bank of Philadelphia, CMU, FGV, UCI, UCLA, UCSB, UIUC, Wisconsin, and the Wharton Macro Lunch. Lian Chen provided expert research assistance. The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. All errors are our own responsibility.

†Yeshiva University Sy Syms School of Business. Email: assac@sas.upenn.edu

‡University of Illinois Urbana-Champaign. Email: kargar@illinois.edu

§Federal Reserve Bank of Philadelphia. Email: benjamin.lester@phil.frb.org

¶UCLA, NBER and CEPR. Email: poweill@econ.ucla.edu

1 Introduction

In many over-the-counter (OTC) markets, dealer banks provide liquidity through their willingness to hold *inventory*: they absorb assets onto their balance sheets when investors need to sell quickly, and they use these assets to fulfill investors' buy orders without delay. After the Global Financial Crisis (GFC) of 2007-2008, several regulations were introduced that increased the cost to dealers of holding inventory.[1] Not surprisingly, dealers responded by reducing their inventory holdings. For instance, according to data from the Flow of Funds, the share of outstanding corporate bonds and non-agency mortgage-backed securities held by broker-dealers fell from 2-3% in 2006 to less than 1% in 2018.[2] At the same time, both market participants and academics alike argued that post-GFC regulations posed a threat to market liquidity.[3] Since maintaining liquid financial markets is crucial for a well-functioning economy, understanding and quantifying the effects of post-GFC regulations on market liquidity and welfare has emerged as a central challenge.

In this paper, we develop a structural model of dealer-intermediated OTC markets in order to meet this challenge. Our starting point is the benchmark search-theoretic framework developed by Duffie, Gârleanu, and Pedersen (2005) and extended to allow for arbitrary preferences and asset holdings by Lagos and Rocheteau (2009) and Gârleanu (2009). However, a key abstraction in these papers is that inventories do not play any economic role for market making. Indeed, in these models, dealers *never* hold inventory—they merely enable customers to access a frictionless market, for which they charge a fee. Our innovation is to bring inventories back into market making with a simple and, arguably, quite natural constraint: we assume that a dealer can only sell to customers the assets that she currently holds in inventory. As a result of this "inventory-in-advance" constraint, a central feature of our model is that dealers choose an optimal amount of inventory in order to

[1]These regulations include the capital requirements embedded in the Dodd-Frank Act; the ban on proprietary trading embedded in the Volcker rule; the incremental risk charge introduced in Basel II.5 (Adrian, Boyarchenko, and Shachar, 2017); and the supplementary leverage ratio (SLR) and liquidity coverage ratio (LCR) that are part of the Basel III framework. However, as Fleckenstein and Longstaff (2020) argue, several other factors increased the funding costs of dealer banks after the GFC, further increasing the costs of balance sheet space; see, e.g., Duffie (2017) and Andersen, Duffie, and Song (2019), who emphasize the role of debt overhang.

[2]Source: Table L.213 of the Federal Reserve's Flow of Funds, shown in Figure 5.

[3]See the extensive discussions by Thakor (2012), Duffie (2017), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018), and the many references therein.

provide liquidity to their customers. Technically, this model is more difficult to analyze than its predecessors because it no longer admits closed form solutions. However, using standard recursive methods, we characterize the equilibrium and study how dealers' optimal inventory holdings depend on various features of the economic environment, such as the frequency and variation in investors' preferences—which captures their trading needs—and the flow (dis)utility dealers receive from holding assets on their balance sheets—which captures the effects of regulatory costs imposed by policymakers. Hence, our model provides a structural framework to evaluate the effects of various regulatory or technological changes in OTC markets on asset prices, transaction costs, trading volume, dealers' profits, and investors' surplus.

We apply this framework to the secondary market for U.S. corporate bonds in order to quantitatively assess the impact of rising inventory costs associated with post-GFC regulations in a specific OTC market. Our quantitative analysis proceeds in two steps. First, using transaction-level data, we calibrate the model to match several target moments constructed from the corporate bond market data before the GFC. Given the parameter values implied by our calibration, we find that, before the GFC, the inventory-in-advance constraint had a relatively modest effect on equilibrium outcomes, relative to an environment where dealers did not face such a constraint. In particular, relative to the frictionless benchmark, we find that the environment without an inventory constraint—but with search and bargaining frictions—creates a welfare loss of 0.87%, while the addition of the inventory constraint increases the welfare loss to 1.25%. Intuitively, the implied cost of holding inventory before the GFC was relatively small, and hence dealers held sufficient inventory to fulfill most customer-buy orders in full.

Second, we study the impact of increasing dealers' balance sheet costs, holding all else equal, on their equilibrium inventory holdings, transaction costs, and welfare. We find that even conservative estimates of the increase in balance sheet costs from post-crisis regulations have significant effects on dealers' equilibrium inventory holdings. For example, if the suite of capital and liquidity requirements introduced after the GFC increased dealers' balance sheet costs by 30 bps—from our pre-crisis estimate of 20 bps to the conservative post-crisis estimate of 50 bps provided by

Fleckenstein and Longstaff (2020)—our model predicts that aggregate dealer inventories would fall by about 40%. If we impose larger estimates of the effects of post-crisis regulations on balance sheet costs, such as the tenfold increase estimated by Du, Tepper, and Verdelhan (2018), our model can account for almost all of the observed decline in dealers' inventory holdings. However, the calibrated model predicts that rising inventory costs have a relatively modest effect on transaction costs; we estimate that post-crisis regulations increase average bid-ask spreads by somewhere between 1 and 4 bps.

We then show that a modest increase in transaction costs does not necessarily imply a modest increase in welfare loss. Indeed, an important advantage of our structural approach is that it allows us to go beyond the analysis of trading costs and generate additional predictions that are difficult to make in reduced-form models. We start by calculating welfare, which is crucial for distinguishing between distributional effects—such as shifts in the share of surplus that accrues to dealers vs. customers—and distortions to the efficient allocation. We find that the welfare cost stemming from the combination of search frictions and inventory constraints increased substantially after the introduction of post-GFC regulations: for the range of post-GFC balance sheet costs we think are plausible, the welfare loss increases to at least 1.75%, and could be as high as 2.4%, which is nearly double the estimated pre-GFC welfare loss. Hence, we find that a reduction in dealers' inventory holdings can have significant consequences for misallocation, despite only modest increases in observed bid-ask spreads. Finally, moving beyond the welfare of investors trading in the corporate bond market, we show that increased inventory costs can also have significant consequences for the price of corporate bonds themselves, and hence on firms' cost of capital.

Related literature

Our theory contributes to the literature that uses search-theoretic models of trade to study OTC markets. Many of these papers build on the basic framework developed in Duffie, Gârleanu, and Pedersen (2005), including the important contributions by Lagos and Rocheteau (2009) and Gârleanu (2009), who extend the benchmark to accommodate arbitrary preferences and asset holdings. Importantly, in most papers in this literature, dealers are assumed to have unfettered access to a frictionless, inter-dealer market, which obviates the need for dealers to hold inventory. Such papers include, but are not limited to, Feldhütter (2012), Lester, Rocheteau, and Weill (2015), Milbradt (2017), Pagnotta and Philippon (2018), Lagos and Zhang (2020), Kargar, Passadore, and Silva (2020), Pinter and Üslü (2021), Palleja (2022), and Li (2023). See Weill (2020) for a thorough review of the literature. Of course, the result that dealers hold no inventories makes these models more tractable, highlighting the role of search and bargaining frictions in the determination of prices and allocations. However, it also makes them ill-suited to study dealers' incentives to hold inventory and provide liquidity in response to various changes in the economic environment and the consequences for asset prices, transaction costs, trade size, volume, and welfare.

In the literature on search-based OTC markets, several papers have proposed models of dealers' inventory management. In Weill (2007) and Lagos, Rocheteau, and Weill (2011), for example, dealers find it optimal to hold inventories in anticipation of aggregate fluctuations in customers' demand. However, in both environments, optimal inventory holdings are always zero in the long run, i.e., in the non-stochastic steady state. In our model, inventories play a non-trivial economic role in the intermediation process— even in the non-stochastic steady state—which we believe is an important feature for studying the long-run decline in inventories between 2008 and 2018.

There are a few contemporaneous papers that develop search-based models of OTC markets in which dealers have non-zero inventory holdings in the steady-state equilibrium. For example, An (2018) studies an environment in which imperfectly competitive dealers have incentive to hold inventories in order to provide immediacy and gain market power with their customers. Tse and Xu (2021) develop a model where dealers are heterogeneous with respect to their inventory capacity in order to rationalize empirical observations about inter-dealer trades in OTC markets. Diao, Dudley, and Sun (2023) allow dealers to choose the intensity with which they contact other market participants, which allows them to study the premium that dealers earn by being more central in the network of dealers.[4] Each of these papers impose simplifying assumptions in order to study

[4]There are also a number of papers in which all agents, including those who play the role of dealers, trade in decentralized markets. See, e.g., Hugonnier, Lester, and Weill (2020, 2022), Shen, Wei, and Yan (2021), Üslü (2019),

specific patterns in the data. For example, the first two papers discussed above restrict the asset holdings of customers to $\{0, 1\}$, whereas the third paper assumes that customers are short-lived, in the sense that they do not incorporate the value of future trading opportunities. Our goal, in contrast, is to develop a more general framework—with forward-looking investors who have arbitrary heterogeneity in both preferences and asset holdings—that can be used to study a variety of empirical observations.[5]

Outside of search-based models, there is also, of course, a celebrated literature on inventory management by dealers, starting with Amihud and Mendelson (1980), Ho and Stoll (1981, 1983), and Mildenstein and Schleef (1983). Relative to this literature, our main contribution is to consider a model in which customers' supply and demand are derived from explicit, dynamic optimization problems, subject to search frictions. This enables us to quantify the gains from trade created by the inter-dealer market and offer a welfare analysis of post-GFC regulations.

Finally, given the focus of our application, our paper is related to several recent empirical studies that have attempted to identify the effect of post-crisis regulations on market liquidity, including Adrian, Boyarchenko, and Shachar (2017), Trebbi and Xiao (2019), Bao, O'Hara, and Zhou (2018), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018), Dick-Nielsen and Rossi (2019), and Choi, Huh, and Shin (2024). By studying this issue within the context of a structural equilibrium model, our analysis complements these studies in several important ways. First, by calibrating our model to match moments before and after the introduction of new regulations, we are able to infer the implicit cost of these regulations on dealers; this cost is difficult to measure directly and, to the

Farboodi, Jarosch, and Shimer (2022), Farboodi, Jarosch, and Menzio (2017), Bethune, Sultanum, and Trachter (2022), Yang and Zeng (2019), and Nosal, Wong, and Wright (2019). In these models, since agents face a short-selling constraint, those who play the role of dealers must hold inventories. Though these models have proven useful in studying the determinants of *inter-dealer* market structure and trading patterns, we assume instead that the inter-dealer market is centralized. This simplification allows us to focus our analysis more squarely on the issue at hand; to derive new, testable implications regarding, e.g., the relationship between dealers' inventory costs and the distribution of trade size.

[5]One advantage of our approach is that it requires a minimal departure from the benchmark model of Lagos and Rocheteau (2009). However, introducing an inventory-in-advance constraint implies that our model no longer admits closed-form solutions for the value functions and distributions. Hence, in addition to our substantive contribution, we also make several methodological contributions, adapting standard recursive methods from Stokey and Lucas (1989) to formally establish key properties of the equilibrium in our environment. See Rocheteau, Weill, and Wong (2018) and Choi and Rocheteau (2021) for related methodological contributions in the New Monetarist literature.

best of our knowledge, such an estimate is new to the literature.[6] Second, while existing empirical studies based on difference-in-difference regressions identify "local" effects of new regulations on a particular measure of liquidity, such as price impact, our model allows us to explore the broader implications of policy for the behavior of customers and dealers, and the subsequent implications for a variety of outcomes, both observable (such as bid-ask spreads, trade size, or volume) and unobservable (such as the time customers wait to complete their trade). Third, and perhaps most important, our structural equilibrium model provides natural measures of welfare, along with the opportunity to perform counterfactuals, which is crucial for evaluating the quantitative impact of policy.[7]

The remainder of the paper has two parts. In Section 2, we describe the model, show that an equilibrium exists, and study analytically a number of its properties. In Section 3, we calibrate the model to the U.S. Corporate Bond market and study the welfare impact of post-GFC regulation.

2 The Model

We consider a continuous time, infinite horizon model of an over-the-counter asset market in the spirit of Gârleanu (2009) and Lagos and Rocheteau (2009). There are two types of infinitely lived agents: a measure of customers normalized to one and a measure $\mu > 0$ of dealers. There is one asset that is durable, perfectly divisible, and in fixed supply, $s > 0$.

We assume that customers have stochastically varying preferences defined over the quantity of assets they hold and a numéraire consumption good. In particular, let $u(q, \delta) + c$ denote a customer's flow utility, where $q \geq 0$ denotes the units of asset the customer holds, δ denotes her current preferences for assets, and c denotes her net consumption (or production if negative)

[6] Allen and Wittwer (2023) develop a model to study dealers' optimal bidding behavior, abstracting from the supply of and demand for assets from the customer sector. They use this framework to estimate the effects of capital constraints—one likely source of the dealer costs that we study—on bond prices and yields.

[7] In this sense, our work is also related to papers that structurally estimate models of OTC markets. Examples of search-based papers include Feldhütter (2012), Gavazza (2016), Brancaccio, Li, and Schürhoff (2017), Hendershott, Li, Livdan, and Schürhoff (2020), Liu (2020), Pinter and Üslü (2021), and Brancaccio and Kang (2022), while examples of network-based papers include Gofman (2014, 2017) and Eisfeldt, Herskovic, Rajan, and Siriwardane (2023). We contribute to this literature by developing a new model and focusing on a different market phenomenon.

of the numéraire good. We assume that $u(q, \delta)$ is strictly increasing and strictly concave in $q > 0$, is continuously differentiable, and satisfies the Inada conditions $\lim_{q \to 0} u_q(q, \delta) = +\infty$ and $\lim_{q \to \infty} u_q(q, \delta) = 0$. We also assume that $u_q(q, \delta)$ is strictly increasing in δ, where u_q denotes the partial derivative with respect to q. Hence a larger preference shock δ creates a stronger demand for the asset.

Preference shocks arrive at rate γ, at which time a new δ' is drawn according to the cumulative distribution function (CDF) $F(\delta')$.[8] We assume that the CDF has support included in some compact interval $[\underline{\delta}, \overline{\delta}]$ but otherwise we make no restrictions; in particular, the CDF can be discrete (as in, e.g., Lagos and Rocheteau, 2009), continuous, or a mixture of the two. For simplicity, we assume that dealers have linear preferences that do not change over time: a dealer receives flow utility $v\,q + c$ from holding q units of the asset in inventory and consuming c units of the numéraire good. All agents discount the future at rate $r > 0$.

Dealers have continuous access to a frictionless, competitive market where they can buy or sell any amount of the asset at price $P > 0$. Customers do not meet each other and trade directly. Instead, customers meet a randomly chosen dealer at independent Poisson arrival times with intensity λ. If there are gains from trade, the two bargain over the terms of trade. We denote by $\theta \in [0, 1]$ the dealers' bargaining power.

Our key departure from the existing literature is an inventory-in-advance constraint: when a dealer meets a customer, she can buy any quantity of assets from the customer, but she can only sell assets that she currently holds in inventories. After completing a transaction, a dealer can then access the inter-dealer market and rebalance her portfolio, either selling the assets she just accumulated or buying assets to restore an optimal level of inventory.

[8]Micro-foundations for such a specification have been provided earlier in the literature. For example, under appropriate specification, $u(q, \delta)$ represents the flow certainty equivalent of holding q units of the asset. See Weill (2020) for a survey.

2.1 Customers

Let $V(q,\delta)$ denote the maximum attainable expected discounted utility of a customer with current asset holdings q and preferences δ. The Hamilton-Jacobi-Bellman (HJB) equation for $V(q,\delta)$ can be written as:

$$rV(q,\delta) = u(q,\delta) + \gamma\mathbb{E}^F\left[V(q,\delta') - V(q,\delta)\right] + \lambda\left[V(q',\delta) - V(\delta,q) - P(q'-q) - \phi\right], \tag{1}$$

where $\mathbb{E}^F[\cdot]$ denotes the expectation with respect to the CDF $F(\delta')$. The interpretation of the HJB equation is standard: the customer enjoys the flow utility $u(q,\delta)$ until one of two events occurs. First, at rate γ, a preference shock arrives, at which time a new δ' is drawn from $F(\delta')$. Second, at rate λ, the customer has the opportunity to trade with a dealer. At this time, the dealer transfers $q'-q$ units of the asset in exchange for the payment $P(q'-q)+\phi$. This payment is comprised of the cost (or revenue) of purchasing (selling) the asset at the inter-dealer price, $P(q'-q)$, plus an intermediation fee, ϕ.

A customer in state (q,δ) and a dealer holding $i \geq 0$ units of the asset choose a pair $(\hat{q},\hat{\phi})$ to maximize the Nash product

$$\left[V(\hat{q},\delta) - V(q,\delta) - P(\hat{q}-q) - \hat{\phi}\right]^{1-\theta}\hat{\phi}^{\theta},$$

subject to the inventory-in-advance constraint

$$0 \leq \hat{q} \leq q + i. \tag{2}$$

Maximizing with respect to $\hat{q}$ reveals that the optimal post-trade asset holding, q', maximizes the trade surplus,

$$q' \in \arg\max \quad V(\hat{q},\delta) - V(q,\delta) - P(\hat{q}-q) \tag{3}$$

subject to (2). Given the value q' that solves this program, the transfer ϕ is set so that the dealer appropriates a fraction θ of the maximized joint surplus:

$$\phi = \theta \left[V(q', \delta) - V(q, \delta) - P\left(q' - q\right)\right]. \tag{4}$$

In what follows, we will adopt the usual convention of using a lower case i to denote an individual dealer's inventory and an upper case I to denote the choice of other dealers. Therefore, in a symmetric, steady-state equilibrium in which $i = I$, substituting (3) and (4) into the HJB equation yields

$$\begin{aligned} rV(q, \delta) &= u(q, \delta) + \gamma \mathbb{E}^F \left[V(q, \delta') - V(q, \delta)\right] \\ &\quad + \lambda(1-\theta) \max_{0 \le q' \le q+I} \left\{ V(q', \delta) - V(q, \delta) - P\left(q' - q\right)\right\}. \end{aligned}$$

Informally differentiating with respect to q and applying the envelope condition yields

$$\begin{aligned} rV_q(q, \delta) &= u_q(q, \delta) + \gamma \mathbb{E}^F \left[V_q(q, \delta') - V_q(q, \delta)\right] \\ &\quad + \lambda(1-\theta) \left[\max\left\{ V_q(q+I, \delta), P\right\} - V_q(\delta, q)\right]. \end{aligned}$$

Let $\Sigma(q, \delta) \equiv V_q(q, \delta) - P$ denote the marginal trade surplus, i.e., the marginal value to a customer of an additional unit of asset, net of the inter-dealer price. We can rewrite the expression above as

$$\begin{aligned} \left[r + \gamma + \lambda(1-\theta)\right] \Sigma(q, \delta) &= u_q(\delta, q) - rP + \gamma \mathbb{E}^F \left[\Sigma(q, \delta')\right] \\ &\quad + \lambda(1-\theta) \max\left\{\Sigma(q+I, \delta), 0\right\}. \end{aligned} \tag{5}$$

Since equation (5) characterizes $\Sigma(q, \delta)$ for any given (P, I), it will sometimes be helpful to make this dependence explicit by writing $\Sigma(q, \delta \mid P, I)$; otherwise, we will suppress this dependence to simplify notation. Notice that the environment of Lagos and Rocheteau (2009), where there is no inventory-in-advance constraint, corresponds to the case where $I \to \infty$ and the final term in (5) disappears. Proposition 1 studies the fixed point equation defined by (5).

Proposition 1. *For all $P > 0$, equation (5) admits a unique, continuous solution $\Sigma(\cdot)$ that has the following properties:*

1. *it is the basis of a solution to the HJB equation (1);*
2. *it is strictly increasing in δ, strictly decreasing in q and P, and weakly decreasing in I;*
3. *for all $\delta \in [\underline{\delta}, \overline{\delta}]$, $\lim_{q \to 0} \Sigma(q, \delta) = \infty$;*
4. *there exists $\hat{q}$ such that, for all $\delta \in [\underline{\delta}, \overline{\delta}]$ and $q > \hat{q}$, $\Sigma(\delta, q) < 0$.*

The first point states that a value function $V(q, \delta)$ solving the HJB equation (1) can be constructed based on $\Sigma(q, \delta)$; the details are in the appendix. Importantly, the construction confirms that the envelope condition that we used informally earlier indeed holds. The properties in the last three points are inherited from the flow marginal value, $u_q(q, \delta) - rP$, except for one: the marginal surplus is decreasing in aggregate inventories, I. Indeed, if I is smaller, customers anticipate that the inventory-in-advance constraint is more likely to bind in the future. This makes it harder for them to accumulate assets, reduces their asset holding and raises their marginal value for the asset.

The function $\Sigma(\cdot)$ entirely characterizes a customer's optimal trading behavior. To see this, note that if the customer and the dealer were unconstrained by inventories, then they would trade to the "target holding" $q^\star(\delta \mid P, I)$ such that the marginal trade surplus is equal to zero, i.e.,

$$\Sigma(q^\star(\delta \mid P, I), \delta \mid P, I) = 0.$$

Proposition 1 ensures that this equation has a unique solution. It also implies some intuitive relationships between an individual customer's current state, the aggregate state, and the customer's target asset holdings. In particular, as one might expect, the customer's target asset position, $q^\star(\delta \mid P, I)$, is increasing in his idiosyncratic preference shock δ and decreasing in the price P. A new feature of our model is induced by the inventory constraints: customers now have incentive to acquire additional assets out of precautionary motives. This incentive grows stronger as I declines.

That is, since an additional unit of the asset is more valuable when dealers hold less inventory, ceteris paribus, $q^\star(\delta \mid P, I)$ is decreasing in I.

2.2 Dealers

Let $\Phi(q,\delta)$ denote the joint distribution of asset holdings and preference shocks across customers. We characterize this distribution below and note for now that optimal trading behavior implies that its support is included in $[0,\overline{q}] \times [\underline{\delta},\overline{\delta}]$, for some $\overline{q} > q^\star(\overline{\delta})$. Using the Nash bargaining solution, we can write the dealer's (flow) profit function as

$$r\Pi(i) = (v - rP)i + \frac{\lambda}{\mu}\theta \int_{(q',\delta')} \max_{0 \le q'' \le q'+i} \left\{V(q'',\delta') - V(q',\delta') - P\left(q''-q'\right)\right\} \, d\Phi(q',\delta'),$$

where we use $\int_{(q',\delta')}$ to denote the integral over $(q',\delta') \in [0,\overline{q}] \times [\underline{\delta},\overline{\delta}]$. Hence, the dealer's objective function has two components: the flow payoff from owning i units of the asset, $(v - rP)i$, and the expected capital gains from trading with a randomly selected customer.

Lemma 1. *For any $\Phi(q,\delta)$, P, and I, the profit function $\Pi(i)$ is concave and continuously differentiable in i, with derivative:*

$$r\frac{d\Pi}{di}(i) = v - rP + \frac{\lambda}{\mu}\theta \int_{(q',\delta')} \max\left\{\Sigma(q'+i,\delta'), 0\right\} \, d\Phi(q',\delta').$$

The expression for the derivative of the profit function is intuitive. The first term is the direct flow utility that a dealer enjoys by holding a marginal unit of the asset. The second term is the user cost: what the dealer has to pay per unit of time to hold a marginal unit of the asset. The third term is the marginal impact of increasing inventory on intermediation profits. Indeed, the dealer meets customers with intensity λ/μ and appropriates a fraction θ of the marginal trading surplus created by increasing inventories, which is equal to $\max\{\Sigma(q'+i,\delta'),0\}$ with a customer of type (q',δ'). Notice in particular that this marginal surplus is strictly positive if and only if $q' + i < q^\star(\delta)$, that is, if and only if it relaxes a binding inventory-in-advance constraint and helps the customer to trade closer to the target.

The first-order condition for the dealer's optimal inventory holdings is simply

$$\Pi'(i) \leq 0, \text{ with equality if } i > 0. \tag{6}$$

Note that a solution to (6) requires $rP \geq v$; if $v > rP$, then dealers would have incentive to acquire infinite inventory. Hence, in equilibrium, the price will adjust to incorporate the dealers' flow value from holding the asset *and* the marginal benefit of increasing inventory on intermediation profits.

As in our analysis of the customer's optimal asset position, the properties of $\Sigma(\cdot)$ allow for some natural, partial equilibrium comparative statics with respect to an individual dealer's optimal inventory holdings. In particular, given the behavior of all other agents (and, hence, aggregate variables), one can show that an individual dealer's optimal i is increasing in the rate at which he meets customers, λ/μ, and the fraction of the trading surplus he extracts through bargaining, θ.[9]

2.3 The steady-state distribution

We now characterize the steady-state distribution $\Phi(q, \delta)$. In a steady state, the gross outflow from any Borel set B of $[0, \overline{q}] \times [\underline{\delta}, \overline{\delta}]$must be equal to the gross inflow:

$$(\gamma + \lambda)\Phi(B) = \int_{(q,\delta)} \left(\gamma \int_{\delta'} \mathbb{I}_{\{(q,\delta') \in B\}} \, dF(\delta') + \lambda \mathbb{I}_{\{(\min\{q^\star(\delta), q+I\}, \delta) \in B\}} \right) d\Phi(q, \delta).$$

The left-hand side is the gross outflow: customers leave the set B when they change utility, with intensity γ, or when they trade, with intensity λ. The right-hand side is the gross inflow. It states that a customer of type (q, δ) may transition into the set B in two ways. First, with intensity γ and probability $dF(\delta')$, she draws the new utility shock δ' and her new type (q, δ') belongs to B.

[9]Keep in mind, however, that these partial equilibrium comparative statics are not necessarily true in general equilibrium. For example, while an increase in the dealers' bargaining power θ provides greater incentive to accumulate inventory, it also changes the distribution of asset holdings across investors in equilibrium and, hence, the size of the trades that investors will desire and their willingness to pay for these trades.

Second, with intensity λ, she receives a trading opportunity, and her new type $(\min\{q^\star(\delta), q+I\}, \delta)$ belongs to B.

Dividing both sides by $(\gamma + \lambda)$ reveals that the steady-state distribution solves the following fixed point problem

$$\Phi = T^\star\left[\Phi\right], \text{ where } T^\star[\Phi](B) = \int_{(q,\delta)} \mathbb{P}(q,\delta,B)\, d\Phi(q,\delta) \tag{7}$$

and

$$\mathbb{P}\left(q,\delta,B\right) = \frac{\gamma}{\lambda+\gamma}\int \mathbb{I}_{\{(q,\delta')\in B\}}\, dF(\delta') + \frac{\lambda}{\lambda+\gamma}\mathbb{I}_{\{(\min\{q^\star(\delta),q+I\},\delta)\in B\}}. \tag{8}$$

The function $\mathbb{P}(q,\delta,B)$ is the transition probability function for the state of a customer when she draws a new preference shock or receives a trading opportunity. After checking appropriate regularity conditions, one can apply Theorems 11.12 and 12.3 in Stokey and Lucas (1989) to establish the following results.

Proposition 2. *Assume that $P > 0$ and $I > 0$. Then, there exists a unique steady-state distribution $\Phi(q,\delta \mid P, I)$. This distribution has the following properties:*

1. *it is decreasing in P in that, for any bounded function $h(q,\delta)$ that is increasing in q, the function $P \mapsto \int h(q,\delta)\, d\Phi(q,\delta \mid P,I)$ is decreasing in P;*
2. *it is weakly continuous in (P, I);*
3. *given any initial condition, Φ_0, the sequence $T^{\star n}[\Phi_0] \to \Phi$ strongly.*

The monotonicity property implies that the law of demand holds in steady state: higher prices are associated with lower aggregate asset holdings by customers. Together with the continuity property, it is crucial to our equilibrium existence proof. The strong convergence result is useful to compute moments since it allows us to calculate any stationary moment by successive iteration.

2.4 Equilibrium

An equilibrium is made up of the following objects: a marginal trade surplus function $\Sigma(q, \delta \mid P, I)$, an optimal inventory holdings of each dealer I, a joint distribution of asset holdings and preferences $\Phi(q, \delta \mid P, I)$, and an inter-dealer price P. These objects must satisfy the following conditions:

(i) $\Sigma(q, \delta \mid P, I)$ solves the Bellman equation (5) given I and P;

(ii) $i = I$ solves the dealer's optimality condition (6) given Φ and P;

(iii) $\Phi(q, \delta \mid P, I)$ is the stationary distribution solving (7);

(iv) the asset market clears

$$\int_{(q',\delta')} q' \, d\Phi(q', \delta' \mid P, I) + \mu I = s. \tag{9}$$

It is easy to construct an equilibrium with $I = 0$ by choosing a ν sufficiently small. In such an equilibrium, there is no trade: when $I = 0$, customers can never buy and, thus, for the market to clear, the equilibrium inter-dealer price must be small enough to ensure that no customer finds it optimal to sell. Hence, in the steady-state the asset is randomly allocated across customers, so that $\Phi(q, \delta) = \Psi(q)F(\delta)$ for some distribution Ψ of asset holdings. Assuming that the support of the distribution of asset holdings has an upper bound $\overline{q}$, this implies that P should be chosen so that the customer with the strongest incentive to sell chooses to hold on to her asset, i.e., $\Sigma(\overline{q}, \underline{\delta} \mid P, 0) \geq 0$. Finally, ν has to be small enough so that, given the distribution Ψ and the marginal trade surplus function, Σ, the dealer's first-order condition (6) holds with inequality.

Next, we characterize the set of ν such that there exists an equilibrium with active intermediation and trade, i.e., with $I > 0$. Clearly, some ν belongs to this set if and only if there exists some $I > 0$ and some $P > 0$ such that the market-clearing condition (9) is satisfied and the dealer's first-order condition (6) holds with equality. Because the stationary distribution is decreasing and weakly continuous in P, it can be shown that the market-clearing condition implies a unique market clearing price given I, denoted by $P(I)$. Plugging this price into the dealer's first-order condition at equality, one obtains that $v = \mathcal{V}(I)$, where

$$\mathcal{V}(I) \equiv rP(I) - \frac{\lambda}{\mu}\theta \int_{(q',\delta')} \max\left\{\Sigma(q'+I,\delta' \mid P(I), I), 0\right\} \, d\Phi(q',\delta' \mid P(I), I). \tag{10}$$

Hence, the set of v such that there exists an equilibrium with $I > 0$ is simply the range of the function $\mathcal{V}(I)$ above, for all values of I that may arise in an equilibrium with active intermediation—in particular, for all $I \in (0, s/\mu)$.

The range of I is unbounded above because $\mathcal{V}(I) \to \infty$ as $I \to s/\mu$. This is true for two reasons. First, when $I \to s/\mu$ customers hold almost no assets, which implies that their marginal utility and the inter-dealer price $P(I)$ go to infinity. Second, since dealers' inventories are bounded away from zero but customers' asset holdings go to zero, the inventory-in-advance constraint never binds. Hence, $\mathcal{V}(I) = rP(I) \to \infty$. In the Appendix, we show that the marginal trade surplus—and, hence, $\mathcal{V}(I)$—is bounded below as $I \to 0$. Together with continuity, this implies that $\mathcal{V}(I)$ remains bounded below over $(0, s/\mu)$. The next theorem summarizes.

Theorem 1. *There exists a $\underline{v}$ such that an equilibrium with active intermediation exists if $v > \underline{v}$ and does not exist if $v < \underline{v}$.*

Importantly, Theorem 1 establishes that our framework provides explicit conditions under which a (fixed) set of dealers finds it optimal to engage in active intermediation. These conditions depend on the arrival rate of trading opportunities, the distribution of preference shocks, the bargaining power of dealers, and the costs that dealers face from holding inventory. Hence, in principle, our framework could be used to understand an important outstanding question in the market microstructure literature—namely, why market makers have successfully entered some markets, but have failed after attempting to enter other (seemingly similar) markets (see Rust and Hall, 2003, for an extended discussion).

As a final note, while we cannot rule out multiple equilibria in this environment, it is easy to study this question numerically. In particular, one sees that multiple equilibria arise whenever there is a region of I where the function $\mathcal{V}(I)$ is decreasing; since $\lim_{I \to s/\mu} \mathcal{V}(I) = +\infty$, the Intermediate Value Theorem implies that there are several I mapping to the same v, which is just another way to say that the same v can be associated with several I.

2.5 Discussion: The inventory constraint and implications

The key assumption in our framework is that dealers can only sell assets that they have acquired before meeting a customer-buyer. In this section, we first discuss this assumption in more detail. Then, we highlight an important qualitative implication of this constraint and provide empirical support from the data.

Inventory in advance. In our model, dealers have to acquire an asset in order to sell it. In real-world markets, one might wonder whether dealers could agree to sell an asset that they do not own, and then either purchase or borrow the asset from another market participant within the allotted amount of time. The key assumption in our framework is that doing so would be costly. For the sake of establishing a simple, tractable benchmark, we assume that it is infinitely costly, i.e., that dealers simply can't sell an asset that they don't own. However, all of the theoretical results are preserved if the additional cost of either borrowing the required quantity of bond on the inter-dealer market or purchasing it outright—as opposed to having the bond ready to sell in advance—is sufficiently large. Several pieces of evidence in the existing literature suggest that such costs are, indeed, significant.

To start, the market for borrowing or "shorting" corporate bonds is also an OTC market, and the costs of shorting a bond can erode most (if not all) of the rents a dealer earns by selling it. For example, Asquith, Au, Covert, and Pathak (2013) estimate that the average cost of shorting a bond is between 10 and 20 bps and can be significantly more for less liquid bonds. Moreover, dealers who short a bond will typically want to hedge the price risk by selling a credit default swap (CDS). This hedging strategy incurs additional costs, such as upfront payments, transaction costs, collateral requirements, and marked-to-market expenses.

Of course, dealers could purchase the promised quantity of bonds outright on the inter-dealer market but this, too, is an OTC market. Therefore, they would have to purchase the bond quickly—which typically requires paying a larger spread—and run the risk of a settlement failure if they cannot locate the necessary quantity within the allotted time. For most of our sample period, "normal" settlement in the corporate bond market had to occur within two days, though this time span was recently shortened to one day. Kargar, Lester, Plante, and Weill (2023) estimate time to trade in the most liquid segment of the corporate bond market and find that a significant fraction of customer inquiries fail to trade immediately and, conditional on such a failure, the average purchase inquiry takes approximately two days to fulfill. Hence, the risk of a settlement failure can be non-trivial, especially for less liquid bonds that are hard to locate quickly. Fleming and Garbade (2005) discuss the various costs associated with settlement fails in the Treasury market, including foregone interest, capital charges, and the reputational costs that derive from an increase in perceived counterparty risk.

Another aspect of our framework that is worth discussing is that we assume it is costly for dealers to hold inventory in advance in order to sell assets to customers, but it is *not* costly for dealers to hold cash in advance in order to buy assets from customers (as in, e.g., Geromichalos, Herrenbrueck, and Lee, 2023). Dealers have a variety of short-term lending vehicles to minimize the opportunity cost of holding cash. However, to the extent that these costs limit dealers' ability to purchase large quantities of bonds from customer-sellers, this restriction could (at least partially) offset some of the asymmetries we discuss in the next section.

Asymmetries implied by the inventory constraint. Before proceeding to our quantitative analysis, we highlight an important qualitative implication of introducing an inventory constraint into an otherwise-standard model of OTC trade. Namely, since the inventory constraint only has a direct impact on purchases, but not on sales, it creates asymmetries that are unique to our model relative to Lagos and Rocheteau (2009).

Consider first a customer who seeks to purchase from a dealer who holds i assets. While she

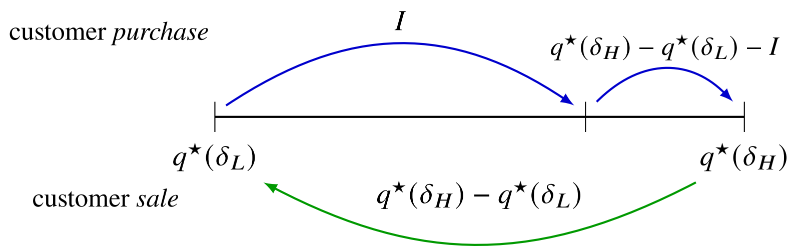

Figure 1. An illustration of the asymmetry between purchases and sales in the special case in which the distribution of preference shocks has a discrete support and can take only two values $\delta_L < \delta_H$, with parameters such that $I < q^\star(\delta_H) - q^\star(\delta_L) < 2I$.

ideally wants to trade up to her target $q^\star(\delta)$, the inventory constraint implies that it is not always feasible. Instead, she will trade to be as close as possible to the target given the constraint; that is, $q' = \min\{q^\star(\delta), q + i\}$. Clearly, sales are never constrained by inventories in this way: a customer who holds $q > q^\star(\delta)$ does not face a constraint on the size of her trade and is able to reach her target in one transaction. Therefore, relative to an otherwise identical trading opportunity without a constraint ($i = \infty$), a customer purchase is smaller but a customer sale is not. Figure 1 illustrates.

A similar asymmetry holds for proportional transaction costs, i.e., for the proportional markup and markdown charged by dealers over the inter-dealer price. Consider first a customer purchase $q^\star(\delta) > q$. From equation 4, with Nash bargaining the trading fee ϕ is a constant share of the surplus. Hence, as a function of *marginal* surplus, the proportional transaction cost is equal to

$$\frac{\theta}{P\left(\min\{q^\star(\delta), q + i\} - q\right)} \int_q^{\min\{q^\star(\delta), q+i\}} \Sigma(x, \delta)\, dx.$$

Clearly, since the marginal surplus is decreasing, the transaction cost above decreases in i. In other words, relative to an otherwise identical trading opportunity without a constraint, a customer-buyer pays a larger transaction cost. A customer-seller, on the other hand, is not constrained by inventories and pays the same transaction cost as she would in the absence of an inventory constraint.

Corollary 1. *Relative to otherwise identical trading opportunities without an inventory constraint, customer-buyers trade smaller quantities and pay larger proportional transaction costs. Customer-sellers trade identical quantities and pay identical proportional transaction costs.*

One may wonder about the empirical implications of this corollary: Does it imply that customer purchases are smaller and more expensive than customer sales? The answer to this question is not obvious since, in principle, the size and the cost of purchases and sales may differ *even without* inventory constraints. However, in Lemma 4 of Appendix B, we establish that, in an otherwise identical model *without* an inventory constraint, purchases and sales have the same average size. Consequently, in the model *with* a constraint, we find numerically that customer purchases are smaller, on average, than customer sales. Since, in the aggregate, the total quantity of assets purchased and sold are equal, this means that our model predicts that the number of customer purchases will be larger than the number of sales. These asymmetries in trade size and the number of trades are well-known empirical features of several major OTC markets (see, for example, Green, Hollifield, and Schürhoff, 2007). Hence, introducing inventory constraints is not just a conceptual contribution, but it also brings the model closer to the data along this important dimension.

Asymmetries in transaction costs, however, can go either way. The reason is that, in the absence of an inventory constraint, proportional transaction costs can be different for purchases than sales. For example, under their preferred parameterization, the model of Duffie, Gârleanu, and Pedersen (2005) implies that transaction costs are zero for purchases and strictly positive for sales. In Appendix B, we show that, in the model of Lagos and Rocheteau (2009) with an isoelastic utility function, $u(q, \delta) = \frac{q^{1-1/\eta}}{1-1/\eta}\delta$, value-weighted transaction costs for purchases are strictly smaller than those for sales if and only if $\eta > 2$. In these cases, we find numerically that the inventory in advance constraint generally raises the transaction cost for purchases, but not always by a sufficient amount to make them larger than the transaction costs for sales.

3 Quantitative Exercise

In this section, we use our model to quantitatively evaluate the market impact of post-GFC regulations that increased dealers' inventory or balance sheet costs. Relative to a purely empirical approach that focuses on traditional liquidity measures, such as transaction costs, our structural approach has the advantage that it allows us to evaluate welfare. For example, it is plausible that post-GFC regulations increased transaction costs without impacting the asset allocation, in which case it would have only had redistributive effects. However, we find the opposite: according to our analysis, post-GFC regulations had significant effects on the allocation of the asset, but endogenous changes in trading behavior left transaction costs unchanged.

Our quantitative analysis proceeds in two steps. First, we calibrate our model to match moments from the corporate bond market before the GFC. Then, to study the impact of rising balance sheet costs, we evaluate the effects of reducing ν on liquidity, prices, allocations, and welfare.

3.1 Data

We use the academic version of the Trade Reporting and Compliance Engine (TRACE) database of U.S. corporate bond transactions, made available by the Financial Industry Regulatory Authority (FINRA). The raw TRACE data provides detailed information on all secondary market transactions self-reported by FINRA member dealers. These include the bond's CUSIP, trade execution time and date, transaction price ($100 = par), the volume traded (in multiple of par), a buy/sell indicator, and flags for dealer-to-customer and inter-dealer trades. Unlike the public version, the academic TRACE does not censor trade volume at $5 million (for investment grade bonds) or $1 million (for high-yield bonds). The academic version also contains masked dealer identities as well as transactions in privately traded Rule 144A bonds that are not disseminated to the public.

Dealers are required to correct errors in previously reported trades with flags corresponding to trade cancellations, modifications, or reversals. We use the standard cleansing algorithm described in Dick-Nielsen (2009, 2014) and Dick-Nielsen and Poulsen (2019) to remove these self-reported errors. Our TRACE sample starts in July 2002 and covers transactions until June 2020. We exclude the COVID-19 crisis period in March and April 2020 from our sample, since our maintained assumption of a non-stochastic steady state is not appropriate for such a turbulent period (see Kargar, Lester, Lindsay, Liu, Weill, and Zúñiga, 2021 for an empirical study).

We collect issue credit ratings and bond characteristics from Mergent Fixed Income Securities

Database (FISD). We drop all bonds not contained in FISD and only consider CUSIPs in TRACE identified by FISD as fixed-coupon U.S. corporate debentures and U.S. corporate bank notes with non-missing maturity dates and amounts outstanding. We also exclude bonds with equity-like and special features.[10] Furthermore, we exclude trades associated with new issuances and also remove transactions that occur within 90 days of the traded bond issuance. This ensures that trading activity in the sample closely aligns with the non-stochastic steady state envisioned by the model.

Our model features a representative asset and a representative customer. Of course, in reality both assets and customers in the corporate bond market are heterogeneous. Hence, we apply two filters to control for heterogeneity while keeping the sample economically relevant. First, we concentrate on trades for investment-grade (IG) bonds, which exhibit more homogeneous liquidity properties and represent the vast majority of daily trading volume in TRACE.[11] Second, since it would be unreasonable to require that our representative customer model rationalizes the significant difference in trade size between institutional and retail investors, we focus on trades larger than $1 million, which are more likely to originate from institutional rather than retail investors.

Finally, in practice, dealers provide liquidity via "agency" and "risky-principal" trades. In an agency trade, the dealer acts as a match maker between buyers and sellers, and never actually owns the asset being traded. In contrast, in a risky-principal trade, dealers buy and sell on their own account, absorbing sell orders onto their balance sheet and fulfilling buy orders by reducing their inventory holdings. Since trading costs for these two types of trades have been shown to be quite different (see e.g., Choi, Huh, and Shin, 2024)—and, in our theory, inventories are a key input into the provision of liquidity services—it is natural for us to focus exclusively on risky-principal transactions. We identify these transactions in the data using the procedure described in Kargar, Lester, Lindsay, Liu, Weill, and Zúñiga (2021).

Table 1 reports summary statistics for the daily number and volume of inter-dealer, customer-

[10]Following earlier work, we exclude all bonds that are convertible, puttable, exchangeable, preferred, asset-backed, secured lease obligations, unit deals, or Yankee bonds. Additionally, we do not consider bonds with variable coupons or sinking funds, or those issued in a foreign currency or as part of unit deals.

[11]In the third quarter of 2023, IG bonds represent approximately 85% of the average total daily trading volume of publicly traded U.S. corporate bonds. Source: U.S. corporate bond statistics from SIFMA.

bought and customer-sold trades for our final, filtered sample. Note that, while the total volume of customer buys and sells are very similar (approximately \$2.2 billion), we observe, on average, more customer buys than customer sells, which is a key qualitative prediction of our model with a (binding) inventory-in-advance constraint.

3.2 Calibration to pre-GFC corporate bond market: Targets

We set the discount factor, r, equal to 5%. We assume that customers have an isoelastic utility function of the form

$$u(q, \delta) = \frac{q^{1-1/\eta}}{1 - 1/\eta}\delta.$$

In addition, we assume that the preference shock, δ, is an iid draw from a discretized log-normal distribution, $F(\delta)$. Given our choice of an isoelastic utility function, the model is homogeneous in s, the per-capita supply of the asset.[12] Hence, we normalize s such that the asset supply held by customers is one, and the aggregate dealer inventories represent 2% of the total asset supply, similar to the dealer sector's pre-GFC corporate bond holding share. Finally, we normalize the mean of $F(\delta)$ such that the price of the asset equals $1/r$ in a Walrasian equilibrium in which the supply held by customers is one, i.e., the same aggregate quantity of asset they hold in our calibration.

The intensity of contact between customers and dealers is typically difficult to calibrate, because the TRACE data does not offer direct evidence about customers' search process. We rely on the work of Kargar et al. (2023), who leverage proprietary data from an electronic trading platform to measure customers' time to trade in the U.S. corporate bond market. Following their estimate, we set λ so that a customer contacts a dealer every 3 days.

Given the assumptions above, there are six parameters that we need to calibrate: the variance of preference shocks, σ_δ^2; the arrival rate of preference shocks, γ; the elasticity parameter for the

[12] Specifically, suppose that we scale the supply by the constant κ, i.e., $\tilde{s} = \kappa s$. Then, scaling preference shocks by the same factor, $\tilde{\delta} = \kappa^{1/\sigma}\delta$, renders the marginal utilities the same for all investors if they scale their holdings by κ. As a result, the equilibrium price remains unchanged, as do the holdings of customers and dealers relative to the aggregate supply.

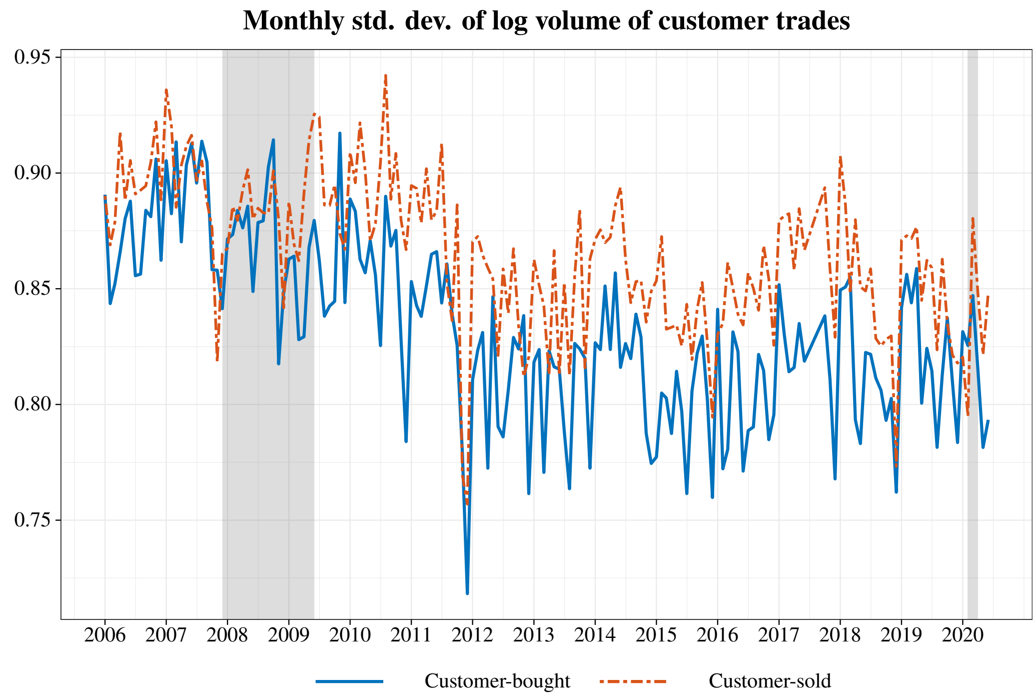

Figure 2. Monthly standard deviation of log trading volume for customer-to-dealer trades in percentage points. We restrict the sample to the subset of trades involving risky-principal trades of investment-grade bonds with size exceeding \$1 million. The vertical shaded bars indicate NBER recessions. Sources: Academic TRACE and FISD.

customers' utility function, η; the dealers' bargaining power, θ; the dealers' utility parameter, ν; and the measure of active dealers, μ. In what follows, we describe the target moments that we use to determine these parameters. Though in general the target moments and parameters are determined simultaneously, we try to connect each moment to the parameter it affects most directly.

The variance of preference shocks. The dispersion in preference shocks determines customers' equilibrium asset holdings and the size of trades they execute when their asset holdings differ from their target portfolios. Hence, to help identify σ_δ^2, we target the standard deviation of log trade size after including dealer fixed effects and controlling for several trade characteristics (such as log amount outstanding, credit rating, time-to-maturity, and coupon rate). In our sample of IG bond trades with par value exceeding \$1 million, we find that the monthly standard deviation of log trade size is about 0.84. Figure 2 plots the monthly standard deviation of log trade size for customer-bought and customer-sold trades.

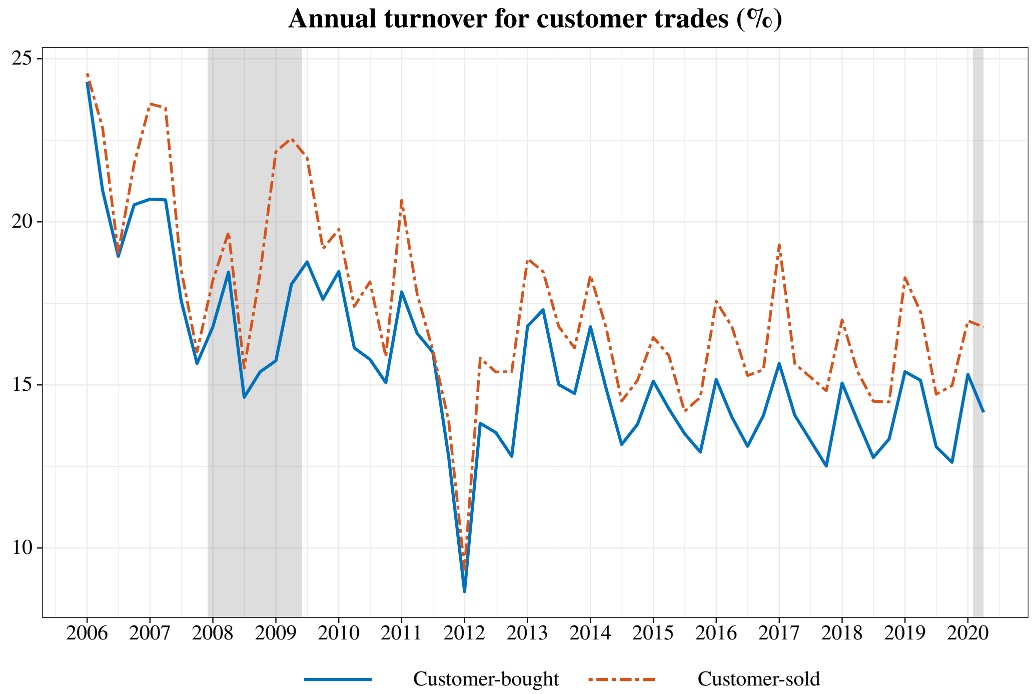

Figure 3. Annual turnover for customer-to-dealer trades in percentage points. We restrict the sample to the subset of trades involving risky-principal trades of investment-grade bonds with size exceeding $1 million. The vertical shaded bars indicate NBER recessions. Sources: Academic TRACE and FISD.

The frequency of preference shock. This parameter is a key determinant of how frequently customers want to buy or sell. Therefore, to help determine γ, we target the turnover of assets that customers purchase, generated by trades greater than $1 million. This is calculated by dividing the total quarterly trading volume by the quarterly average of the amount outstanding of bonds for dealer-to-customer transactions in our sample. We find that the turnover is approximately 20% annually. Figure 3 plots the annual turnover for customer-bought and customer-sold trades.

Dealers' bargaining power and the elasticity of customers' utility function. To help determine these parameters we focus on proportional transaction costs. On the one hand, dealers' bargaining power determines the overall level of transaction costs paid by customers. On the other hand, the elasticity of the customer's utility function is one determinant of the asymmetry between transaction costs for customer purchases and customer sales. This is demonstrated formally in Lemma 5 of Appendix B; it shows that, without an inventory-in-advance constraint, a lower value of η tends to make the transaction costs for customer purchases larger than that for sales.

Now turning to measurement, our empirical target is based on the value-weighted two-way trading cost proposed by Choi, Huh, and Shin (2024):

$$2Q \times \frac{\text{traded price} - \text{reference price}}{\text{reference price}}, \tag{11}$$

where Q is equal to +1 for a customer *buy* and −1 for a customer *sell*. For each customer trade, a "reference price" is calculated as the volume-weighted average price of inter-dealer trades larger than \$100,000 in the same bond and on the same day, excluding inter-dealer trades executed within 15 minutes. The measure is calculated at the trade level for all customer trades classified as risky-principal, and then calculated at the bond-day level by taking the volume-weighted average of trade level spreads.

We find that, for risky-principal customer trades involving IG bonds with size exceeding \$1 million, two-way trading costs from Choi, Huh, and Shin (2024) for buy and sell transactions in the pre-GFC periods are 10.8 bps and 9.6 bps, respectively. Figure 4 plots the volume-weighted average round-trip (two-way) customer-bought and customer-sold transactions costs from Choi, Huh, and Shin (2024).

Measure and utility flow of dealers. The measure of dealers μ and their utility flow v jointly determine how much inventory they hold on aggregate, $\mu \times I$. Moreover, dealers' individual choice of inventory, I, determines how many purchases customers need to make in order to reach their target holding. Hence, we propose two targets that help determine μ and v: the total asset holdings of the dealer sector before the GFC, as a share of outstanding assets; and the ratio of the number of customer-sell transactions to the number of customer-buy transactions.

To set a target for the asset holdings of the dealer sector prior to GFC, we rely on data from the Federal Reserve's Flow of Funds. Figure 5 plots the share of corporate and foreign bonds held by security brokers and dealers from the Federal Reserve's Flow of Funds. Up until 2002, this share was slightly above 2%, then it dramatically increased during the years leading up to the GFC, only to drop sharply to levels below 1% after the GFC. The substantial increase leading up to the GFC

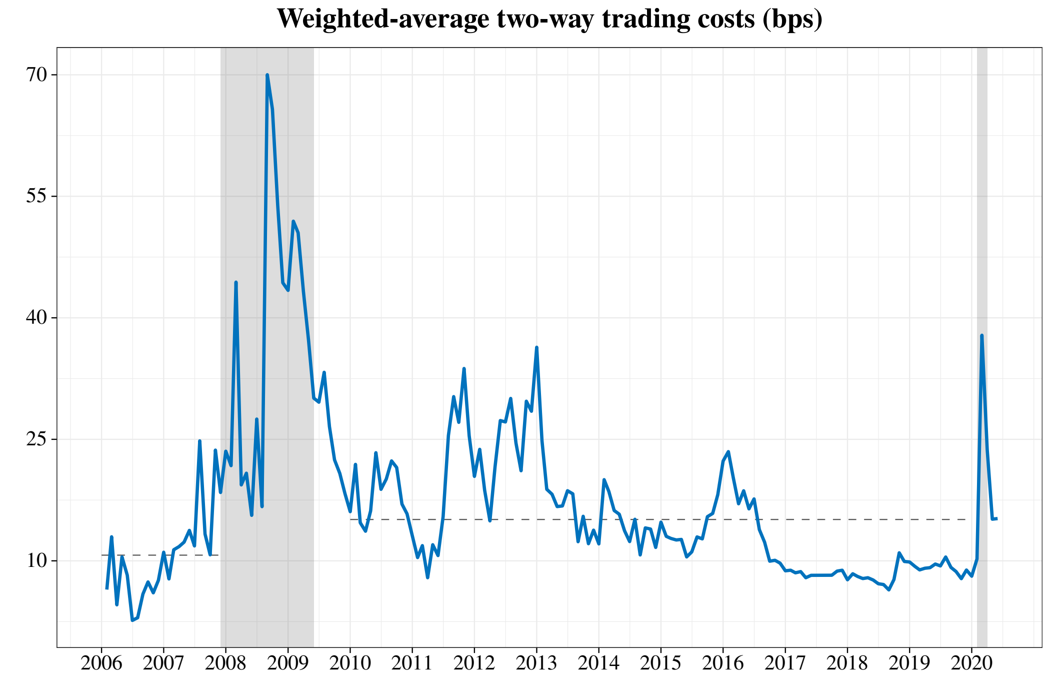

Figure 4. Monthly size-weighted average of customer-bought and customer-sold two-way (round-trip) trading costs proposed by Choi, Huh, and Shin (2024) from equation (11). We restrict the sample to the subset of trades involving risky-principal trades of investment-grade bonds with size exceeding \$1 million. The horizontal dashed lines represent subsample averages for the pre-GFC (2006–2007) and post-GFC (2010–2019) periods. The vertical shaded bars indicate NBER recessions. Sources: Academic TRACE and FISD.

may be partly attributed to non-agency mortgage-backed securities (MBS), which are included in the Flow of Funds accounting but not relevant to our quantitative exercise. For this reason, for the pre-GFC calibration, we take the dealer sector asset holding to be 2% of the aggregate asset supply.

In Figure 6, we plot the ratio of the number of customer-sold trades to the number of customer-bought trades in our sample, after adjusting for order flow imbalance (see Appendix C). We calculate that, prior to the GFC, this ratio averaged about 0.76.

3.3 Calibration to pre-GFC corporate bond market: Outcomes

Table 2 presents the target moments and Table 3 reports our calibrated parameters. All pre-GFC targets are matched nearly exactly. Since the Walrasian price of the asset would be $1/r$ and dealers are risk neutral, it is natural to interpret the flow utility that dealers receive from holding a unit of the asset as $v = 1 - \tau$, where τ denotes the (flow) inventory cost to dealers of holding the asset on

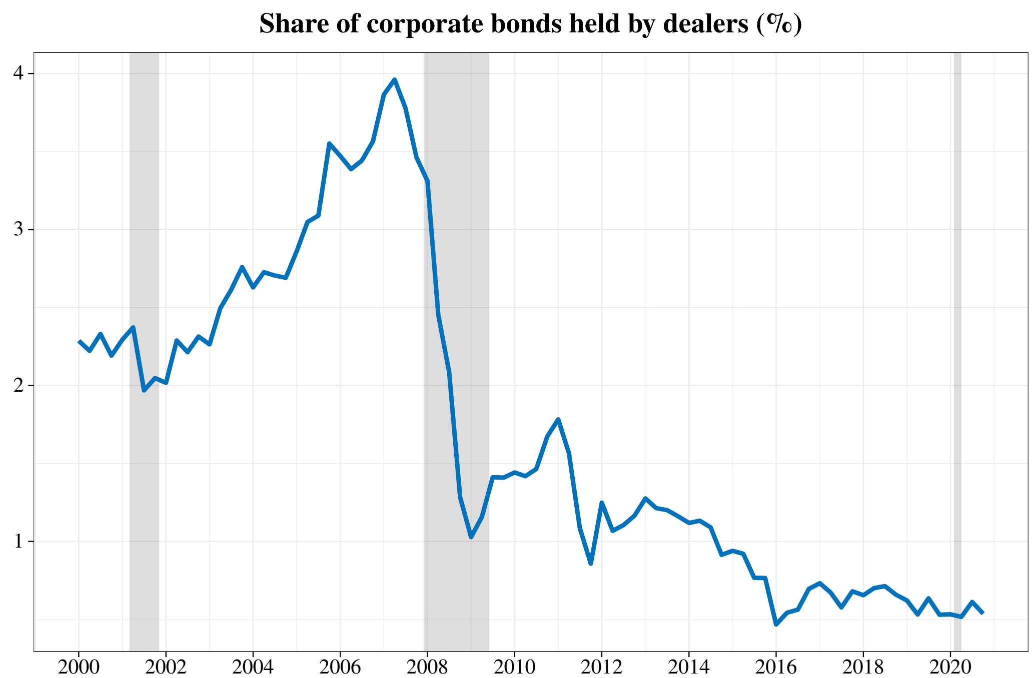

Figure 5. Share of corporate and foreign bond holdings for security broker-dealers. The vertical shaded bars indicate NBER recessions. Source: Table L.213 of the Federal Reserve's Z.1: Financial Accounts of the United States (the Flow of Funds).

their balance sheet. Hence, if we think of the asset as a consol bond, our calibration implies that balance sheet costs before the GFC were approximately 4% of the bond's coupon or, equivalently, about 20 bps (since the average coupon rate of investment grade bonds during our sample period is approximately 5%). Note that our estimate of 20 bps is similar to estimates in the literature; for example, despite using entirely different methods of inference, Fleckenstein and Longstaff (2020) estimate average balance sheet costs of approximately 28 bps in 2006 (see Table 3), whereas Du, Tepper, and Verdelhan (2018) estimate that capital charges associated with balance sheet constraints were approximately 22 bps in 2006 (see Table II in the Internet Appendix).

Figure 7 plots the (log of) customers' target asset holdings, $q^{\star}(\delta)$, along with each dealers' equilibrium inventory holdings, I. To highlight the effects of the inventory constraint in our framework, we also plot the target asset holdings in an environment without the inventory constraint (i.e., the environment of Lagos and Rocheteau, 2009, with the same parameters, assuming that the asset supply is equal to one). The dashed horizontal line is the inventory holdings of dealers. Hence, the inventory constraint only binds for those customers who receive a sufficiently large preference

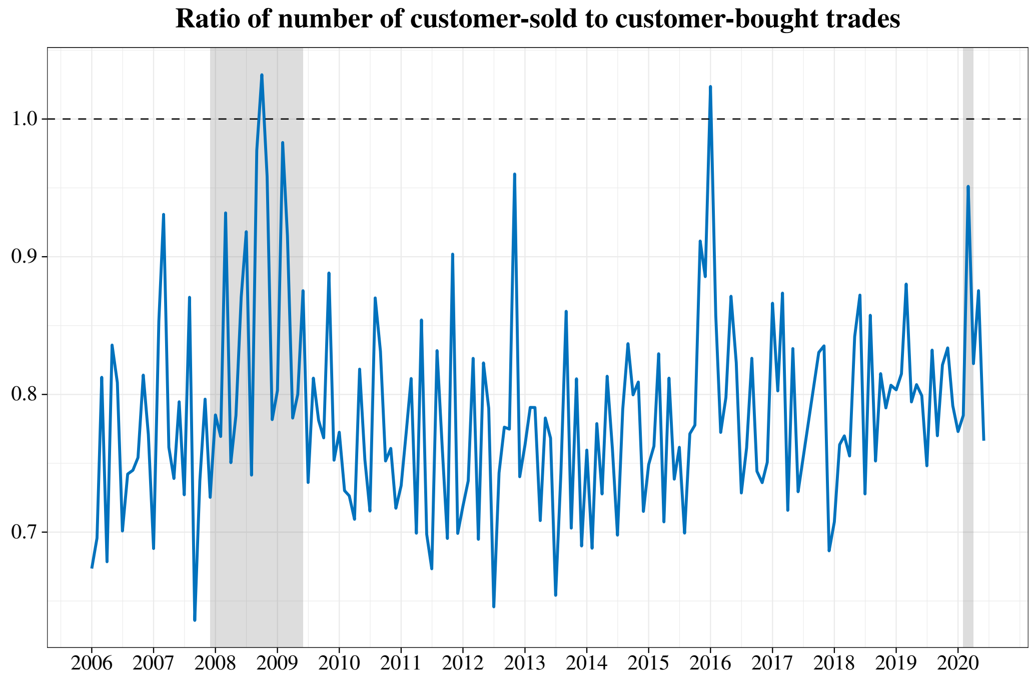

Figure 6. Ratio of the monthly number of customer-sold to customer-bought trades. We adjust this ratio for order imbalance, as described in Appendix C. We restrict the sample to the subset of trades involving risky-principal trades of investment-grade bonds with size exceeding $1 million. The horizontal dashed line at 1 represents the case in which the number of customer-bought trades equals the number of customer-sold trades, as is the case in a model without the inventory constraint. The vertical shaded bars indicate NBER recessions. Sources: Academic TRACE and FISD.

shock. Also note that, at the scale of the figure, the target asset holdings are indistinguishable from the targets in the equilibrium without an inventory constraint (they are in fact slightly larger).

Introducing an inventory constraint creates a small increase in the bid-ask spread charged by dealers: given the parameter values that emerge from our calibration, the trading costs in the no-constraint environment would be approximately 0.5 bps smaller. As trading costs rise, the customers' valuation for the asset declines, which puts downward pressure on the inter-dealer price. However, the presence of an inventory constraint also puts upward pressure on the price, because precautionary incentives increase customers' demand for the asset. In equilibrium, we find that the forces putting downward pressure on the price dominate, as the inter-dealer price in our benchmark model is slightly lower than in the model without an inventory constraint.[13]

[13]The inventory constraint also has an effect on the calibrated value of η, the elasticity of the customer's utility function. In particular, had we ignored the inventory constraint, the imputed value of η would have to be strictly less than 2 in order to generate the observed asymmetry in transaction costs; we prove this formally in Appendix B.4. In contrast, with an inventory constraint, the imputed value of η is strictly greater than 2 and closer to existing estimates in the literature (such as, e.g., Lester, Rocheteau, and Weill, 2015).

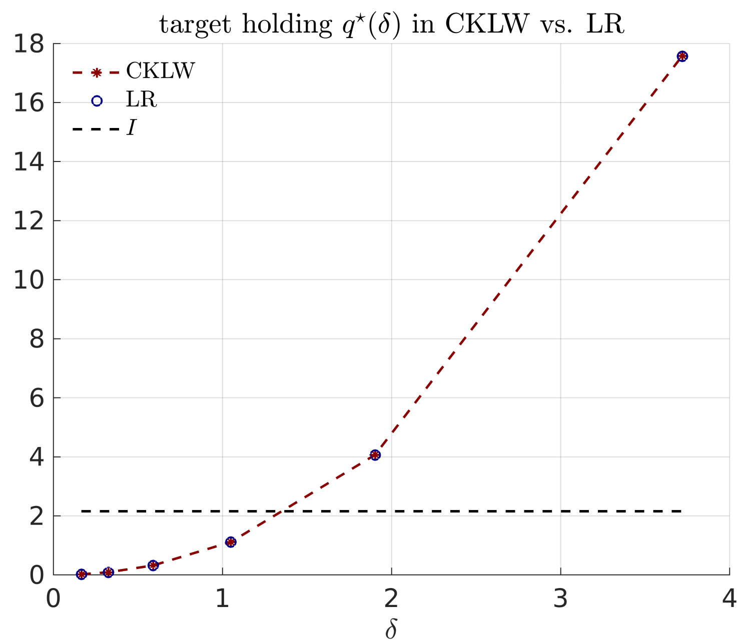

Figure 7. Target asset holdings, $q^\star(\delta)$, with inventory constraints (CKLW) and without (LR). The horizontal line is I in CKLW.

Overall, however, we find that the presence of an inventory constraint in the pre-GFC economy had quite mild effects on equilibrium prices and target holdings, relative to an environment where dealers are not required to hold inventory in order to intermediate trade. To get a sense of the welfare cost of the inventory constraint, we calculate the gains from trade that are realized in equilibrium relative to the gains from trade in a frictionless environment. More precisely, we consider the following measure of welfare *loss*:

$$L = \frac{W_{fb} - W_{eqm}}{W_{fb} - W_{aut}}, \tag{12}$$

where W_{fb}, W_{eqm}, and W_{aut} denote total welfare in the (first best) frictionless environment, in equilibrium, and in autarky, respectively. This measures the fraction of gains from trade lost by the market in equilibrium relative to the first best. As a point of reference, we find that the environment without an inventory constraint—but with search and bargaining frictions—creates a welfare loss of 0.87%. In our environment, where dealers must hold inventory in order to sell, the welfare loss is 1.25%.

3.4 The effects of rising inventory costs

We now study the impact of increasing dealers' costs of holding assets on their balance sheets, holding all else equal. Of course, other aspects of the U.S. corporate bond market have also changed since the GFC. For example, there has been a significant shift in the composition of corporate bond owners towards mutual funds and ETFs, which tend to exhibit different trading behavior than other institutional investors. Another important development in the corporate bond market is the emergence of electronic trading platforms, which have gained market share relative to voice-based trading and have arguably made the market more competitive (O'Hara and Zhou, 2021). However, our intention is not to study the impact of *all* changes in the corporate bond market since the GFC, but rather to isolate the impact of post-crisis regulatory constraints. To that end, we fix all other parameter values at their pre-GFC levels and study the effects of increasing dealers' costs of holding inventory (i.e., reducing v).

Given the many changes that have occurred since the GFC, it is difficult to empirically identify the precise increase in balance sheet costs associated with post-crisis regulations. Fleckenstein and Longstaff (2020) estimate an increase of approximately 32 bps that can be attributed to capital regulation, which would imply an increase in inventory costs in our model from approximately 4% before the GFC to approximately 10% during the post-crisis period. Du, Tepper, and Verdelhan (2018) find a much larger increase in post-GFC regulatory costs: they estimate that capital charges associated with balance sheet constraints increased tenfold between 2006 and the period 2010-2016. In what follows, we don't take a firm stance on the exact increase in balance sheet costs, but rather report the effects of reducing v on inventory holdings, transaction costs, and welfare for a wide range of values.

Inventory holdings. Figure 8 plots dealers' aggregate inventories as a fraction of total supply,

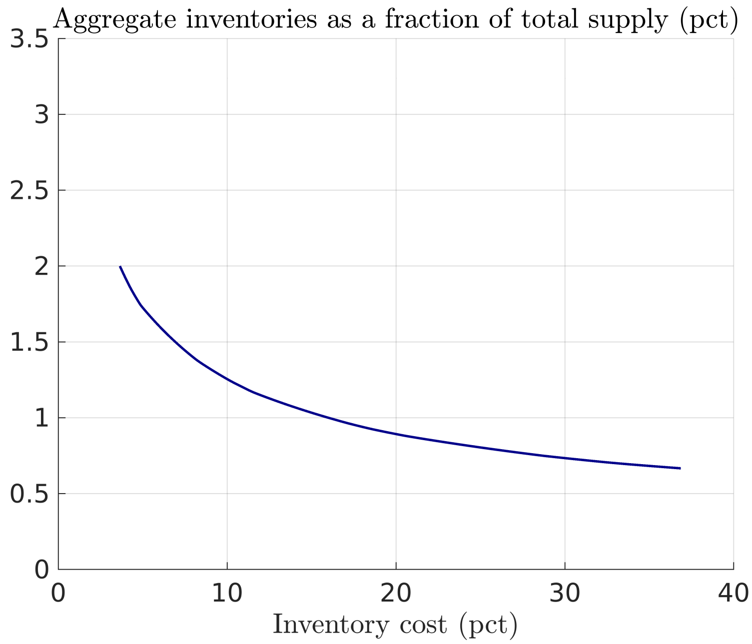

Figure 8. Dealers' aggregate inventory as a percentage of total supply

$\mu I/s$, as a function of the implicit inventory cost, τ. It shows that inventory holdings are relatively sensitive to balance sheet costs: even a more conservative estimate of the increase in balance sheet costs, from 4% to 10% (Fleckenstein and Longstaff, 2020), would generate a decline in dealers' aggregate inventory holdings of approximately 40%. In order to engineer *all* of the observed decline in dealers' inventory holdings illustrated in Figure 5—from 2% to approximately 0.66%—inventory costs would need to increase tenfold, from 4% in the pre-crisis calibration to approximately 40%, as a fraction of the asset coupon, holding all other parameters fixed.

Trading costs. Figure 9 shows the effects of rising inventory costs on trading costs. As one can see, our model predicts that an increase in inventory costs has a relatively modest effect on average spreads: even a tenfold increase in inventory costs, from 4% to 40%, would only generate an increase in the value-weighted trading cost of about 4 bps, which is approximately 80% of the increase we observed in the data. The figure also reveals a striking asymmetry: since rising balance sheet costs imply an ever-tightening inventory-in-advance constraint, trading costs increase

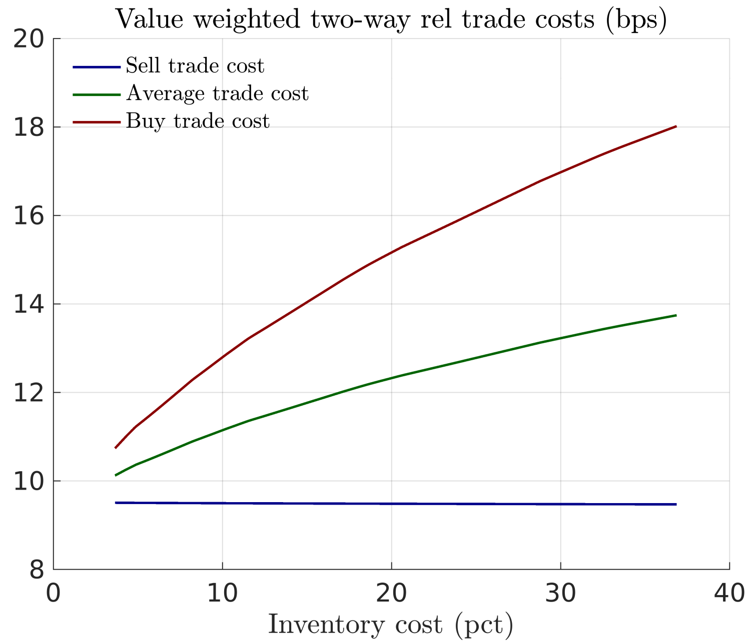

Figure 9. Value weighted trading costs

for customer purchases but not for customer sales, as explained in Section 2.5. In the data, we observe a similar asymmetry, although it is much less pronounced than in the model: two-way transaction costs increased by 5.4 bps for customer purchases and by 4.8 bps for customer sales.

Welfare. Figure 10 illustrates the effects of increasing balance sheet costs on welfare: we plot the welfare loss, defined in (12), as the inventory cost rises. A key insight from our calibration is that a decline in dealers' willingness to hold inventory and provide liquidity can have relatively large effects on welfare *even though the observed effect on transaction costs can be quite small.* For example, an increase in τ from 4% to 10% increases the welfare loss by nearly 50%, even though average transaction costs only increase by 1 bps (or approximately 10%). Intuitively, this finding highlights that a reduction in dealers' inventory holdings can have significant consequences for misallocation—as customers' asset holdings shift further from their frictionless counterpart—despite small increases in observed bid-ask spreads. One important caveat of this calculation, of course, is that it abstracts from the potential benefits that derive from greater financial stability.

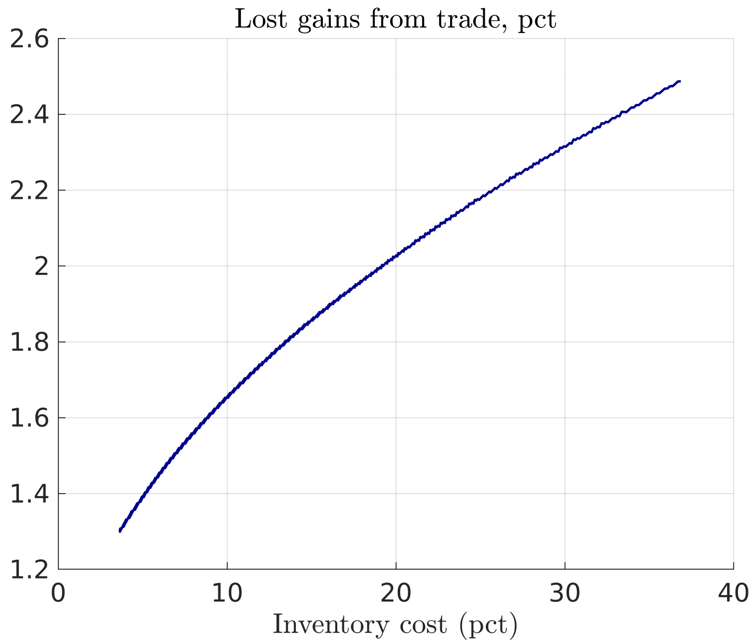

Figure 10. Equilibrium welfare loss

Asset prices and the liquidity yield spread. One may also argue that our measure of welfare is too narrow because it ignores the general equilibrium effects of post-GFC policy on asset prices, and hence on the (unmodeled) firms' cost of capital. One way to measure the impact on the firms' cost of capital is to calculate the liquidity yield spread of corporate bonds implied by our model. Recall that, given the normalization of preference shocks explained above, the frictionless price is equal to $1/r$, the present value of a riskless consol bond with a coupon equal to 1. Hence, it is natural to define the liquidity yield spread based on the following pricing condition: the present value of this consol bond, at rate equal to r plus the liquidity yield spread, should be equal to P, the price of the consol bond in our theoretical OTC market. As a result, the liquidity yield spread can be defined:

$$\text{liquidity yield spread} = \frac{1}{P} - r.$$

Figure 11 plots the relationship between inventory costs and the yield spread. In our pre-GFC

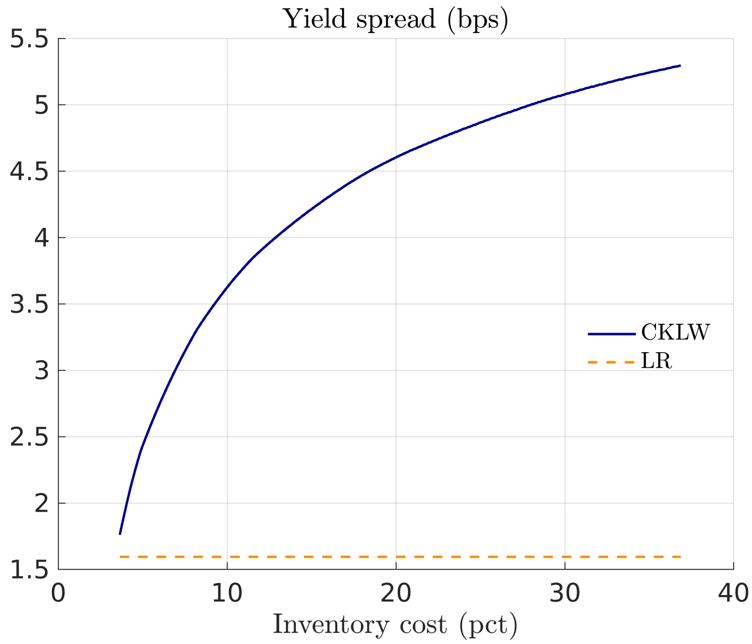

Figure 11. Yield spread with inventory in advance constraint (CKLW) and without (LR)

baseline, the liquidity yield spread is less than 2 bps. Given the range of estimates of τ, our model would predict that the liquidity yield spread doubled, at the very least, as a result of increased regulatory costs. Hence, according to our calibration, the liquidity component of firms' cost of capital would rise substantially as a consequence of an increase in balance sheet costs.

4 Conclusion

We extend the standard search-theoretic model of dealer-intermediated OTC markets, in which dealers never hold inventory, by introducing a simple and natural inventory-in-advance constraint, which makes inventory a necessary input for intermediation. We characterize the equilibrium and study how dealers' optimal inventory choices depend on inventory costs. We calibrate the model to transaction-level data from the corporate bond market and analyze the welfare impact of rising inventory costs associated with post-crisis regulations. We measure the welfare loss as the fraction of total gains from trade that the OTC market fails to generate. Our results indicate that increased balance sheet costs can have a significant impact on welfare, even if observed transaction costs only increase modestly.

Given the tractability of this framework, a number of additional exercises could potentially be fruitful. On the theoretical front, one might like to explore the efficiency properties of the model, which are not straightforward. In particular, since dealers pay the cost to acquire inventories before a match is formed, the well-known "hold up" problem emerges as long as investors extract some of the match surplus (i.e., if $\theta < 1$). In general, this implies that the equilibrium level of investment does not coincide with that of the planner's solution. On the quantitative front, one would naturally like to incorporate other developments in the corporate bond market, and to explore the interaction between these developments and changes in the costs of holding inventory associated with post-crisis regulations. We leave these exercises, and more, for future work.

References

Adrian, Tobias, Nina Boyarchenko, and Or Shachar, 2017, Dealer balance sheets and bond liquidity provision, *Journal of Monetary Economics* 89, 92–109.

Allen, Jason, and Milena Wittwer, 2023, Market power and capital constraints, Working paper, Boston College.

Amihud, Yakov, and Haim Mendelson, 1980, Dealership market: Marketmaking with inventory, *Journal of Financial Economics* 8, 31–53.

An, Yu, 2018, Competing with inventory in dealership markets, Working paper, Johns Hopkins University.

Andersen, Leif, Darrell Duffie, and Yang Song, 2019, Funding value adjustments, *Journal of Finance* 74, 145–192.

Asquith, Paul, Andrea S. Au, Thomas Covert, and Parag A. Pathak, 2013, The market for borrowing corporate bonds, *Journal of Financial Economics* 107, 155–182.

Bao, Jack, Maureen O'Hara, and Alex Zhou, 2018, The Volcker Rule and making in times of stress, *Journal of Financial Economics* 130, 95–113.

Bessembinder, Hendrik, Stacey Jacobsen, William Maxwell, and Kumar Venkataraman, 2018, Capital commitment and illiquidity in corporate bonds, *Journal of Finance* 73, 1615–1661.

Bethune, Zachary, Bruno Sultanum, and Nicholas Trachter, 2022, An information-based theory of financial intermediation, *Review of Economic Studies* 89, 2381–2444.

Bogachev, Vladimir I., 2007, *Measure Theory* (Springer-Verlag).

Brancaccio, Giulia, and Karam Kang, 2022, Search frictions and product design in the municipal bond market, Working Paper 30775, National Bureau of Economic Research.

Brancaccio, Giulia, Dan Li, and Norman Schürhoff, 2017, Learning by trading: The case of the U.S. market for municipal bonds, Working paper, New York University.

Carter, Michael, and Bruce Van Brunt, 2000, *The Lebesgue-Stieltjes Integral: A Practical Introduction* (Springer-Verlag, New York).

Choi, Jaewon, Yesol Huh, and Sean Seunghun Shin, 2024, Customer liquidity provision: Implications for corporate bond transaction costs, *Management Science* 70, 187–206.

Choi, Michael, and Guillaume Rocheteau, 2021, New monetarism in continuous time: Methods and applications, *Economic Journal* 131, 658–696.

Diao, Chengjie, Evan Dudley, and Hongfei Amy Sun, 2023, Search and inventory in the over-the-counter market, Working paper, Queen's University.

Dick-Nielsen, Jens, 2009, Liquidity biases in TRACE, *Journal of Fixed Income* 19, 43–55.

Dick-Nielsen, Jens, 2014, How to clean enhanced TRACE data, Working paper, CBS.

Dick-Nielsen, Jens, and Thomas Kjær Poulsen, 2019, How to clean academic TRACE data, Working paper, CBS and BI Norwegian.

Dick-Nielsen, Jens, and Marco Rossi, 2019, The cost of immediacy for corporate bonds, *Review of Financial Studies* 32, 1–41.

Du, Wenxin, Alexander Tepper, and Adrien Verdelhan, 2018, Deviations from covered interest rate parity, *Journal of Finance* 73, 915–957.

Duffie, Darrell, 2017, Post-crisis bank regulations and financial market liquidity, Banca d'Italia, Thirteenth Paolo Baffi Lecture on Money and Finance.

Duffie, Darrell, Nicolae Gârleanu, and Lasse H. Pedersen, 2005, Over-the-counter markets, *Econometrica* 73, 1815–1847.

Eisfeldt, Andrea L., Bernard Herskovic, Sriram Rajan, and Emil Siriwardane, 2023, OTC intermediaries, *Review of Financial Studies* 36, 615–677.

Farboodi, Maryam, Gregor Jarosch, and Guido Menzio, 2017, Intermediation as rent extraction, Working Paper 24171, National Bureau of Economic Research.

Farboodi, Maryam, Gregor Jarosch, and Robert Shimer, 2022, The emergence of market structure, *The Review of Economic Studies* 90, 261–292.

Feldhütter, Peter, 2012, The same bond at different prices: Identifying search frictions and selling pressures, *Review of Financial Studies* 25, 1155–1206.

Fleckenstein, Matthias, and Francis A. Longstaff, 2020, Renting balance sheet space: Intermediary balance sheet rental costs and the valuation of derivatives, *Review of Financial Studies* 33, 5051–5091.

Fleming, Michael J., and Kenneth Garbade, 2005, Explaining settlement fails, *Current Issues in Economics and Finance* 11, Federal Reserve Bank of New York.

Folland, Gerald B., 1999, *Real Analysis: Modern Techniques and Their Applications* (John Wiley & Sons Inc.).

Gârleanu, Nicolae, 2009, Portfolio choice and pricing in illiquid markets, *Journal of Economic Theory* 144, 532–564.

Gavazza, Alessandro, 2016, An empirical equilibrium model of a decentralized asset market, *Econometrica* 84, 1755–1798.

Geromichalos, Athanasios, Lucas Herrenbrueck, and Sukjoon Lee, 2023, The strategic determination of the supply of liquid assets, *Review of Economic Dynamics* 49, 1–36.

Gofman, Michael, 2014, A network-based analysis of over-the-counter markets, Working paper, University of Rochester.

Gofman, Michael, 2017, Efficiency and stability of a financial architecture with too-interconnected-to-fail institutions, *Journal of Financial Economics* 124, 113–14.

Green, Richard, Burton Hollifield, and Norman Schürhoff, 2007, Dealer intermediation and price behavior in the aftermarket for new bond issues, *Journal of Financial Economics* 86, 643–682.

Hendershott, Terrence, Dan Li, Dmitry Livdan, and Norman Schürhoff, 2020, Relationship trading in over-the-counter markets, *Journal of Finance* 75, 683–734.

Ho, Thomas, and Hans R. Stoll, 1981, Optimal dealer pricing under trading transactions and return uncertainty, *Journal of Financial Economics* 9, 47–73.

Ho, Thomas, and Hans R. Stoll, 1983, The dynamics of dealer markets under competition, *Journal of Finance* 38, 1053–1074.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill, 2020, Frictional intermediation in over-the-counter markets, *Review of Economic Studies* 87, 1432–1469.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill, 2022, Heterogeneity in decentralized asset markets, *Theoretical Economics* 17, 1313–1356.

Kargar, Mahyar, Benjamin Lester, David Lindsay, Shuo Liu, Pierre-Olivier Weill, and Diego Zúñiga, 2021, Corporate bond liquidity during the COVID-19 crisis, *Review of Financial Studies* 34, 5352–5401.

Kargar, Mahyar, Benjamin Lester, Sébastien Plante, and Pierre-Olivier Weill, 2023, Sequential search for corporate bonds, Working Paper 31904, National Bureau of Economic Research.

Kargar, Mahyar, Juan Passadore, and Dejanir Silva, 2020, Liquidity and risk in OTC markets: A theory of asset pricing and portfolio flows, Working paper, UIUC.

Lagos, Ricardo, and Guillaume Rocheteau, 2009, Liquidity in asset markets with search frictions, *Econometrica* 77, 403–426.

Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill, 2011, Crises and liquidity in over-the-counter markets, *Journal of Economic Theory* 146, 2169–2205.

Lagos, Ricardo, and Shengxing Zhang, 2020, Turnover liquidity and the transmission of monetary policy, *American Economic Review* 110, 1635–1672.

Lester, Benjamin, Guillaume Rocheteau, and Pierre-Olivier Weill, 2015, Competing for order flow in OTC markets, *Journal of Money, Credit and Banking* 47, 77–126.

Li, Jessica S., 2023, Strategic bargaining and portfolio choice in intermediated markets, Working paper, University of Chicago.

Liu, Shuo, 2020, Dealer's search intensity in U.S. corporate bond markets, Working paper, Tsinghua University.

Milbradt, Konstantin, 2017, Asset heterogeneity in over-the-counter markets, Working paper, Kellogg School of Management.

Mildenstein, Eckart, and Harold Schleef, 1983, The optimal pricing policy of a monopolistic marketmaker in equity market, *Journal of Finance* 38, 218–231.

Nosal, Ed, Yuet-Yee Wong, and Randall Wright, 2019, Intermediation in markets for goods and markets for assets, *Journal of Economic Theory* 183, 876–906.

O'Hara, Maureen, and Alex Zhou, 2021, Anatomy of a liquidity crisis: corporate bonds in the covid-19 crisis, *Journal of Financial Economics* 142, 46–68.

Pagnotta, Emiliano S., and Thomas Philippon, 2018, Competing on speed, *Econometrica* 86, 1067–1115.

Palleja, Mariano J., 2022, Error filter for the TRACE database, Working paper, UCLA.

Pinter, Gabor, and Semih Üslü, 2021, Comparing search and intermediation frictions across markets, Working paper, Bank of England and Johns Hopkins University.

Rocheteau, Guillaume, Pierre-Olivier Weill, and Tsz-Nga Wong, 2018, A tractable model of monetary exchange with ex post heterogeneity, *Theoretical Economics* 13, 1369–1423.

Rust, John, and George Hall, 2003, Middlemen versus market makers: A theory of competitive exchange, *Journal of Political Economy* 111, pp. 353–403.

Shen, Ji, Bin Wei, and Hongjun Yan, 2021, Financial intermediation chains in an over-the-counter market, *Management Science* 67, 4623–4642.

Stokey, Nancy L., and Robert E. Lucas, 1989, *Recursive Methods in Economic Dynamics* (Harvard University Press, Cambridge).

Thakor, Anjan V., 2012, The economic consequences of the Volcker Rule, Report by the U.S. Chamber of Commerce Center for Capital Market Competitiveness.

Trebbi, Francesco, and Kairong Xiao, 2019, Regulation and market liquidity, *Management Science* 65, 1949–1968.

Tse, Chung-Yi, and Yujing Xu, 2021, Inter-dealer trades in OTC markets–who buys and who sells?, *Review of Economic Dynamics* 39, 220–257.

Üslü, Semih, 2019, Pricing and liquidity in decentralized asset markets, *Econometrica* 87, 2079–2140.

Weill, Pierre-Olivier, 2007, Leaning against the wind, *Review of Economic Studies* 74, 1329–1354.

Weill, Pierre-Olivier, 2020, The search theory of over-the-counter markets, *Annual Review of Economics* 12, 747–773.

Yang, Ming, and Yao Zeng, 2019, Coordination and fragility in liquidity provision, Working paper, UCL and Wharton.

Tables

Table 1. Summary statistics. This table provides mean, standard deviation, 25th, 75th, 50th, and 95th percentiles of the average daily number of trades and volume by counterparty type, all years. The "daily num." variables refer to the daily number of trades and the "daily vol." variables refer to the average total daily volume, in millions USD. "customer" trades refer to trades between a dealer and a customer which represent the sum of "customer-bought" and "customer-sold" trades. The sample is from the academic version of TRACE and runs from July 2002 to the end of June 2020. Our sample only includes trades for investment-grade bonds with size exceeding $1 million and excludes the COVID-19 crisis period in March and April 2020. All agency transactions where dealers act as match makers have been removed. Rule 144A bonds for which trades are not disseminated to the public are excluded. We filter the sample as described in the main text.

Variable	Mean	Std. dev.	Q25	Q50	Q75	Q95
Daily num. inter-dealer	844.40	469.91	584	817.50	1,048	1,663.30
Daily num. customer	988.94	522.80	681	965	1,252.75	1,866.65
Daily num. customer-bought	511.40	272.60	353	497	646	985.55
Daily num. customer-sold	477.54	257.75	322.25	462	614	907.20
Daily vol. interdealer ($m)	2,603.91	1,568.39	1,699.70	2,455.07	3,224.50	5,442.98
Daily vol. customer ($m)	4,516.78	2,389.89	3,154.19	4,384.17	5,697.99	8,484.16
Daily vol. customer-bought ($m)	2,200.03	1,173.49	1,539.42	2,145.60	2,737.46	4,165.64
Daily vol. customer-sold ($m)	2,316.75	1,255.36	1,591.49	2,234.40	2,979.39	4,429.07

Table 2. Calibration targets.
This table presents moments from the TRACE data, which serve as calibration targets for the model. The pre-GFC period spans from 2006 to 2007, while the post-GFC period covers 2010 to 2019. We use the values from the pre-GFC period as our calibration targets.

Moment	Pre-GFC	Post-GFC
Num. customer-sold/num. customer-bought trades	0.7621	0.7794
Two-way spread, customer-sold (bps)	9.5160	13.6105
Two-way spread, customer-bought (bps)	10.8928	16.9578
Two-way spread for customer trades (bps)	10.1066	15.1040
Monthly std. dev. of log trade size, customer-bought	0.8183	0.7779
Monthly std. dev. of log trade size, customer-sold	0.8403	0.8098
Annual turnover, customer-bought (%)	19.9163	14.5082
Annual turnover, customer-sold (%)	21.2388	16.2388

Table 3. Values of calibrated parameters.
This table reports parameter values used in calibrating the model with associated empirical targets in the TRACE data. The target values are reported in Table 2.

	Parameter	Value	Target (target value)
σ_δ^2	Dispersion in preference shocks	0.2183	Std. dev. of log trade size (0.84)
θ	Dealers' bargaining power	0.6493	Avg. two-way customer trading cost, buy (10.8 bps)
η	Elasticity of customers' utility	2.1945	Avg. two-way customer trading cost, sell (9.5 bps)
γ	Preference shock intensity	0.4137	Annual turnover for customer trades (20%)
v	Flow utility of dealers	0.9635	Dealer sector's pre-GFC bond holding share (2%)
μ	Measure of dealers	0.0095	No. customer-sold / No. customer-bought (0.76)

Appendix

A Omitted Proofs

A.1 Proof of Proposition 1

Existence, uniqueness, and continuity. Let $x = (\delta, q, P, I)$ and $X = [\underline{\delta}, \overline{\delta}] \times (0,\infty) \times (0,\infty) \times [0,\infty)$. For any strictly positive q' and P', let $X' = [\underline{\delta}, \overline{\delta}] \times [q',\infty) \times (0, P'] \times [0,\infty)$. Now consider the set $C_b(X')$ of bounded continuous functions of $x \in X'$, equipped with the sup norm. For any $h \in C_b(X')$, define the operator:

$$\begin{aligned} T[h](q,\delta \mid P, I) =& \frac{u_q(q,\delta) - rP}{r+\gamma+\lambda(1-\theta)} + \frac{\gamma}{r+\gamma+\lambda(1-\theta)} \mathbb{E}^F\left[h(q,\delta' \mid P, I)\right] \\ &+ \frac{\lambda(1-\theta)}{r+\gamma+\lambda(1-\theta)} \max\{h(q+I,\delta \mid P, I), 0\}. \end{aligned}$$

Since the first term $u_q(q,\delta) - rP$ is bounded on the domain X', so is $T[h]$. Moreover, since $T[h]$ is the sum of continuous functions, it is also continuous. Hence, the operator T maps $C_b(X')$ into itself. Next, one can easily verify that T satisfies Blackwell's sufficient conditions for a contraction (see Theorem 3.3 in Stokey and Lucas, 1989, henceforth SLP), with modulus of contraction $(\gamma+\lambda(1-\theta))/(r+\gamma+\lambda(1-\theta))$. An application of the Contraction Mapping Theorem (see Theorem 3.2 in SLP) establishes uniqueness of a bounded and continuous solution over any X'. Given uniqueness, this solution can be extended uniquely over the entire set X by letting $q' \to 0$ and $P' \to \infty$.

Conversely, if we consider any solution of the HJB defined over the domain X, then its restriction over the domain X' satisfies the HJB as well and so must coincide with the solution we constructed above.

Monotonicity. The operator T preserves the following weak monotonicity properties: if h is increasing in δ and decreasing in (q, P, I), then so is $T[h]$. Since weak monotonicity properties are preserved by passing to the limit, they are inherited by the fixed point. Now note that the first term of $T[h]$, $u_q(\delta, q) - rP$ is in fact strictly increasing in δ and strictly decreasing in (q, P). Hence, the fixed point, $\Sigma = T[\Sigma]$, also has these strict monotonicity properties.

$\Sigma(q,\delta)$ goes to infinity as $q \to 0$. Given that the third term of the Bellman equation is non-negative, it follows that, for any q,

$$\Sigma(q,\underline{\delta}) \geq \frac{u_q(q,\underline{\delta}) - rP}{r+\gamma+\lambda(1-\theta)} + \frac{\gamma}{r+\gamma+\lambda(1-\theta)} \Sigma(q,\underline{\delta}) \Rightarrow \Sigma(q,\underline{\delta}) \geq \frac{u_q(q,\underline{\delta}) - rP}{r+\lambda(1-\theta)}, \tag{13}$$

where we omitted the dependence of Σ on (P, I) for notational convenience. Since the utility function satisfies Inada conditions, $\lim_{q\to 0} \Sigma(q,\underline{\delta}) = +\infty$. Since $\Sigma(q,\delta)$ is increasing in δ, it follows that $\lim_{q\to 0} \Sigma(q,\delta) = +\infty$ for all δ as well.

$\Sigma(q,\delta) < 0$ for q sufficiently large. Let $\hat{q}$ denote the solution of $u_q(\overline{\delta}, \hat{q}) = rP$. Evaluating $T[\Sigma]$ at $(\overline{\delta}, \hat{q})$—and keeping in mind that Σ is a fixed point—we obtain

$$\Sigma(\hat{q}, \overline{\delta}) = \frac{\gamma}{r+\gamma+\lambda(1-\theta)} \mathbb{E}^F \left[\Sigma(\hat{q}, \overline{\delta}) \right] + \frac{\lambda(1-\theta)}{r+\gamma+\lambda(1-\theta)} \max\{\Sigma(\hat{q}+I, \overline{\delta}), 0\}.$$

Given that Σ is increasing in δ and $\Sigma(\hat{q}+I, \overline{\delta}) \leq \Sigma(\hat{q}, \overline{\delta})$, we obtain

$$\Sigma(\hat{q}, \overline{\delta}) \leq \frac{\gamma}{r+\gamma+\lambda(1-\theta)} \Sigma(\hat{q}, \overline{\delta}) + \frac{\lambda(1-\theta)}{r+\gamma+\lambda(1-\theta)} \max\{\Sigma(\hat{q}, \overline{\delta}), 0\},$$

implying that $\Sigma(\hat{q}, \overline{\delta}) \leq 0$. Given that Σ is strictly decreasing in q and strictly increasing in δ, the result follows.

Using $\Sigma(q,\delta)$ to construct the value function $V(q,\delta)$. The properties established in the previous paragraph implies that the equation $\Sigma(q,\delta) = 0$ has a unique solution, which we denote by $q^\star(\delta)$. Fix some $q_0 \in (0,\infty)$ and let $\delta \mapsto V(q_0,\delta)$ denote the solution to

$$rV(q_0,\delta) = u(q_0,\delta) + \gamma \mathbb{E}^F \left[V(q_0,\delta') - V(q_0,\delta) \right] + \lambda(1-\theta) \int_{q_0}^{\min\{q^\star(\delta), q_0+I\}} \Sigma(x,\delta)\, dx.$$

The existence and uniqueness of such a function is guaranteed by standard contraction-mapping arguments. Our guess for the value function at any (δ, q) is:

$$V(q,\delta) = V(q_0,\delta) + \int_{q_0}^{q} \Sigma(x,\delta)\, dx + P(q-q_0). \tag{14}$$

Note that, since $\Sigma(q,\delta)$ is strictly decreasing, it follows that $V(q,\delta)$ is strictly concave. Next, we verify that $V(q,\delta)$ constructed above solves the HJB equation (1). Namely, multiplying the above equation by r and substituting in the HJB equation for $V(q_0,\delta)$ and $\Sigma(x,\delta)$, we have:

$$\begin{aligned} rV(q,\delta) =& u(q_0,\delta) + \gamma \mathbb{E}^F \left[V(q_0,\delta') - V(q_0,\delta) \right] + \lambda(1-\theta) \int_{q_0}^{\min\{q^\star(\delta), q_0+I\}} \Sigma(x,\delta)\, dx \\ &+ \int_{q}^{q_0} \left(u_q(x,\delta) - rP \right) dx + \gamma \int_{q_0}^{q} \mathbb{E}^F \left[\Sigma(x,\delta') - \Sigma(x,\delta) \right] dx \\ &+ \lambda(1-\theta) \int_{q_0}^{q} \left[\Sigma(\min\{q^\star(\delta), x+I\}, \delta) - \Sigma(x,\delta) \right] dx + rP(q-q_0). \end{aligned}$$

Adding the first term on the first line, the first term on the second line, and the third term on the third line, we obtain $u(q_0,\delta)$, that is, the first term on the right-hand side of the HJB equation (1). Adding the second term on the first line together with the second term on the second line, we obtain $\gamma \mathbb{E}^F \left[V(q,\delta') - V(q,\delta) \right]$, that is, the second term on the right-hand side of the HJB equation (1). Grouping the last two other terms together, we obtain:

$$
\begin{aligned}
&\lambda(1-\theta)\left[\int_{q_0}^{\min\{q^\star(\delta),q_0+I\}} \Sigma(x,\delta)\,dx + \int_{q_0}^{q} \left[\Sigma(\min\{q^\star(\delta),x+I\},\delta) - \Sigma(x,\delta)\right]\,dx\right] \\
=&\lambda(1-\theta)\left[\int_{q_0}^{\min\{q^\star(\delta),q_0+I\}} \Sigma(x,\delta)\,dx + \int_{q_0+I}^{q+I} \Sigma(\min\{q^\star(\delta),x\},\delta)\,dx + \int_{q}^{q_0} \Sigma(x,\delta)\,dx\right] \\
=&\lambda(1-\theta)\left[\int_{q_0}^{\min\{q^\star(\delta),q_0+I\}} \Sigma(x,\delta)\,dx + \int_{\min\{q^\star(\delta),q_0+I\}}^{\min\{q^\star(\delta),q+I\}} \Sigma(x,\delta)\,dx + \int_{q}^{q_0} \Sigma(x,\delta)\,dx\right] \\
=&\lambda(1-\theta)\int_{q}^{\min\{q^\star(\delta),q+I\}} \Sigma(x,\delta) = \lambda(1-\theta)\left(V(q',\delta) - V(q,\delta) - P(q'-q)\right),
\end{aligned}
$$

where $q' \equiv \min\{q^\star(\delta), q+I\}$. In the above, the second line obtains by change of variable, and the third line because, by definition of $q^\star(\delta)$, $\Sigma(\min\{q^\star(\delta),x\}) = 0$ for all $x \geq q^\star(\delta)$. The fourth line follows by piecing the three integrals together and using our definition of $V(q,\delta)$ in equation (14). The last step is to verify that q' maximizes surplus subject to $0 \leq q' \leq q+I$, which follows immediately since $V(q',\delta)$ is strictly concave.

A.2 Proof of Lemma 1

The integrand in the dealer's profit function is the maximized trade surplus:

$$
\begin{aligned}
g(i,q',\delta') &= \max_{0\leq q''\leq q'+i} \{V(q'',\delta') - V(q',\delta') - P\,(q''-q')\} \\
&= V(\min\{q^\star(\delta'),q'+i\},\delta') - V(q',\delta') - P\left(\min\{q^\star(\delta'),q'+i\} - q'\right),
\end{aligned}
$$

where the second equality follows from Section 2.1, where we established that the optimum is attained for $q'' = \min\{q^\star(\delta), q'+i\}$. Since, by its construction in Section 2.1, the value function is continuously differentiable in q', a direct calculation reveals that g is continuously differentiable in i with derivative

$$
0 \leq g_i(i,q',\delta') = \Sigma(\min\{q^\star(\delta),q'+i\},\delta') = \max\{\Sigma(q'+i,\delta'),0\} \leq \max\{\Sigma(q',\delta'),0\}.
$$

The inequality on the right-hand side shows that $|g_i|$ is bounded by an integrable function of (q',δ'). Hence, an application of Theorem 2.27 in Folland (1999) shows that

$$
\int_{(q',\delta')} g(i,q',\delta')\,d\Phi(q',\delta')
$$

is differentiable with respect to i and that its derivative is obtained by differentiating under the integral sign. The result follows.

A.3 Proof of Proposition 2

For this proof, pick some $0 < \underline{P} < \overline{P}$ and $0 < \underline{I} < \overline{I}$. Now, for all $(P, I) \in [\underline{P}, \overline{P}] \times [\underline{I}, \overline{I}]$, define the transition probability function (8) over $[0, \overline{q}] \times [\underline{\delta}, \overline{\delta}]$ where $\overline{q} > q^\star(\overline{\delta} \,|\, \underline{I}, \underline{P})$. One sees that $\overline{q}$ has been chosen sufficiently large so that, for all $(P, I) \in [\underline{P}, \overline{P}] \times [\underline{I}, \overline{I}]$, the stationary distribution will belong to $[0, \overline{q}] \times [\underline{\delta}, \overline{\delta}]$.

Existence, uniqueness, and strong convergence. We rely on Lemma 11.11 and Theorem 11.12 in SLP, which provide sufficient conditions for the existence of the operator $T^{\star N}$ to be a contraction mapping for some N and guarantee that the desired properties hold. Let $\mathbb{P}^N(q, \delta, B)$ denote the probability of reaching the Borel set B in N applications of the transition function when starting in state (q, δ). The sufficient condition, labeled "condition M" by SLP, is that there exists some $\varepsilon > 0$ and some integer N such that, for any Borel set $B \subseteq [0, \overline{q}] \times [\underline{\delta}, \overline{\delta}]$, $\mathbb{P}^N(q, \delta, B) \geq \varepsilon$ for all (q, δ) or $\mathbb{P}^N(q, \delta, B^c) \geq \varepsilon$ for all (q, δ). In our setting condition M follows because, through their trades, customers are always able to reach their target holdings in a uniformly bounded number of trades, so that they eventually transition to the "diagonal" set

$$D \equiv \{(q^\star(\delta), \delta) \,:\, \delta \in [\underline{\delta}, \overline{\delta}]\}. \tag{15}$$

Specifically, consider any Borel set B of $[0, \overline{q}] \times [\underline{\delta}, \overline{\delta}]$ and pick N such that $(N-1)I \geq q^\star(\overline{\delta})$, so a customer reaches her target holding in at most $N-1$ successive trades with dealers. Then we have that, for all (q, δ),

$$\mathbb{P}^N(q, \delta, B) \geq \mathbb{P}^N(q, \delta, B \cap D) \geq \frac{\gamma}{\gamma + \lambda} F\left((B \cap D)_\delta\right) \left(\frac{\lambda}{\gamma + \lambda}\right)^{N-1},$$

where we use the shorthand A_δ to denote the set of δ such that $(q, \delta) \in A$ for some q (the "δ-section" of the set A). In words, the above inequality states that the probability of reaching B in N transitions is greater than the probability of reaching the intersection of B with the diagonal set, which is itself greater than the probability of first drawing a preference shock in the δ-section of $B \cap D$, and receiving $N-1$ successive trading opportunities, which is sufficient to reach a point in $B \cap D$. Likewise, for all (q, δ),

$$\mathbb{P}^N(q, \delta, B^c) \geq \mathbb{P}^N(q, \delta, B^c \cap D) \geq \frac{\gamma}{\gamma + \lambda} F\left([\underline{\delta}, \overline{\delta}] \backslash (B \cap D)_\delta\right) \left(\frac{\lambda}{\gamma + \lambda}\right)^{N-1},$$

since $B^c \cap D = D \backslash (B \cap D)$. Since either $F\left((B \cap D)_\delta\right) \geq 1/2$ or $F\left([\underline{\delta}, \overline{\delta}] \backslash (B \cap D)_\delta\right) \geq 1/2$, condition M holds for

$$\varepsilon = \frac{1}{2} \frac{\gamma}{\gamma + \lambda} \left(\frac{\lambda}{\gamma + \lambda}\right)^{N-1}.$$

Weak continuity with respect to (P, I). One obtains weak continuity with respect to $(P, I) \in [\underline{P}, \overline{P}] \times [\underline{I}, \overline{I}]$ by an application of Theorem 12.13 in SLP. The first condition of the theorem is that the state space is compact, which is true here by assumption. The second condition is that the transition probability function is weakly continuous in (q, δ, P, I). Note that $q^\star(\delta \mid P, I)$ is continuous in (δ, P, I) since it uniquely solves $\Sigma(q, \delta \mid P, I) = 0$, where $\Sigma(q, \delta \mid P, I)$ is continuous in (q, δ, P, I). Hence, the condition follows from Theorem 12.3 in SLP. The third condition is the that, for all (P, I), the operator $T^\star$ has a unique fixed point, which we established in the previous paragraph.

Monotonicity in *P*. Consider, for any bounded and measurable function, g, the conditional expectation operator:

$$T[g](q, \delta \mid P, I) = \int_{(q', \delta')} g(q', \delta' \mid P, I)\, \mathbb{P}\,(q, \delta, dq', d\delta' \mid P, I)$$
$$= \frac{\gamma}{\gamma + \lambda} \int_{\delta'} g(q, \delta' \mid P, I)\, dF(\delta') + \frac{\lambda}{\gamma + \lambda} g\big(\min\{q^\star(\delta \mid P, I), q + I\}, \delta \mid P, I\big).$$

Since $q^\star(\delta \mid P, I)$ is decreasing in P, one sees that the operator preserves the following joint monotonicity property: for any bounded measurable function $g(q, \delta \mid P, I)$ that is increasing in q and decreasing in P, then $T[g](q, \delta \mid P, I)$ is also increasing in q and decreasing in P. By induction, it follows that this is also true for the n-transitions-ahead conditional expectation: $T^n[g](q, \delta \mid P, I)$ is increasing in q and decreasing in P as well. In particular, if $P' \geq P$, then

$$T^n[g](q, \delta \mid P', I) \leq T^n[g](q, \delta \mid P, I).$$

Given the strong convergence result established before, we can pass to the limit and obtain

$$\int g(q', \delta' \mid P', I)\, d\Phi(q', \delta' \mid P', I) \leq \int g(q', \delta' \mid P, I)\, d\Phi(q', \delta' \mid P, I),$$

as claimed.

A.4 Proof of Theorem 1

Lower and upper bounds on target holdings. Equation (13) evaluated at $q^\star(\underline{\delta})$ implies that $u_q(q^\star(\underline{\delta}), \underline{\delta}) - rP \leq 0$, from which it follows that, for all $\delta \in [\underline{\delta}, \overline{\delta}]$,

$$q^\star(\delta) \geq u_q^{-1}\left(rP, \underline{\delta}\right). \tag{16}$$

Now consider the Bellman equation for $\Sigma(q, \overline{\delta})$ evaluated at $q \geq q^\star(\overline{\delta})$ so that $\max\{\Sigma(q + I, \overline{\delta}), 0\} = 0$:

$$\Sigma(q, \overline{\delta}) \leq \frac{u_q(q, \overline{\delta}) - rP}{r + \gamma + \lambda(1-\theta)} + \frac{\gamma}{r + \gamma + \lambda(1-\theta)} \Sigma(q, \overline{\delta}) \implies \Sigma(q, \overline{\delta}) \leq \frac{u_q(q, \overline{\delta}) - rP}{r + \lambda(1-\theta)}.$$

Letting $q \downarrow q^\star(\overline{\delta})$ we obtain that $u_q(q^\star(\overline{\delta}), \overline{\delta}) - rP \geq 0$, implying that for all $\delta \in [\underline{\delta}, \overline{\delta}]$

$$q^\star(\delta) \leq u_q^{-1}\left(rP, \overline{\delta}\right). \tag{17}$$

Market-clearing given inventory. We now establish that, given some $I \in (0, s/\mu)$, there is a unique price $P(I)$ such that the market-clearing condition (9) holds. First, since the stationary distribution of asset holdings $\Phi(q', \delta' \mid P, I)$ is weakly continuous and decreasing in P, it follows that the left-hand side of the market-clearing condition is continuous and decreasing in P as well. Second, the lower and the upper bounds of equations (16) and (17) imply that

$$u_q^{-1}\left(rP, \underline{\delta}\right) \leq \int_{(q',\delta')} q' \, d\Phi(q', \delta' \mid P, I) \leq u_q^{-1}\left(rP, \overline{\delta}\right).$$

Together with the Inada condition, this means that the quantity of assets held by customers goes to infinity as $P \to 0$ and to zero as $P \to \infty$. Hence, an application of the Intermediate Value Theorem implies that the market clearing equation (9) has at least one solution.

To establish uniqueness we show that the left-hand side of (9) is a strictly decreasing function of P. To do so, first note that, by stationarity, trading between customers and dealers keeps the quantity of assets held by the customer sector constant:

$$\int_{(q',\delta')} q' \, d\Phi(q', \delta' \mid P, I) = \int_{(q',\delta')} \min\{q^\star(\delta' \mid P, I), q' + I\} \, d\Phi(q', \delta' \mid P, I). \tag{18}$$

Hence, for any two prices $0 < P_1 < P_2$,

$$\begin{aligned}
&\int_{(q',\delta')} q' \, d\Phi(q', \delta' \mid P_1, I) - \int_{(q',\delta')} q' \, d\Phi(q', \delta' \mid P_2, I) \\
&= \int_{(q',\delta')} \min\{q^\star(\delta' \mid P_1, I), q' + I\} \, d\Phi(q', \delta' \mid P_1, I) - \int_{(q',\delta')} \min\{q^\star(\delta' \mid P_2, I), q' + I\} \, d\Phi(q', \delta' \mid P_2, I) \\
&\geq \int_{(q',\delta')} \Big(\min\{q^\star(\delta' \mid P_1, I), q' + I\} - \min\{q^\star(\delta' \mid P_2, I), q' + I\} \Big) \, d\Phi(q', \delta' \mid P_1, I) \\
&\geq \int_{(q',\delta') \in D} \Big(\min\{q^\star(\delta' \mid P_1, I), q' + I\} - \min\{q^\star(\delta' \mid P_2, I), q' + I\} \Big) \, d\Phi(q', \delta' \mid P_1, I) \\
&= \int_{(q',\delta') \in D} \Big(q^\star(\delta \,|\, P_1, I) - q^\star(\delta \mid P_2, I) \Big) \, d\Phi(q', \delta' \mid P_1, I) > 0,
\end{aligned}$$

where the equality on the second line follows from (18); the inequality on the third line follows from the fact that the stationary distribution is decreasing in P; the inequality on the fourth line follows because target holdings are decreasing in P so that the integrand is positive; and the equality on the last line follows because, on the diagonal set D defined in (15), $(q', \delta') = (q^\star(\delta'), \delta')$. Finally, the strict inequality on the last line follows because target holdings are strictly decreasing in P and the diagonal set D has strictly positive measure. Indeed, for any N such that $NI \geq q^\star(\overline{\delta})$, it takes at most N consecutive trading opportunities to reach the diagonal set from any (q', δ') in the support of the stationary distribution. Hence $\mathbb{P}(q', \delta', D) \geq (\lambda/(\gamma + \lambda))^N$. Now using stationarity we have that

$$\Phi(D) = T^\star [\Phi] (D) = \int_{(q',\delta')} \mathbb{P}(q', \delta', D) \, d\Phi(q', \delta') \geq \left(\frac{\lambda}{\lambda + \gamma}\right)^N > 0.$$

Taken together, we obtain that the market-clearing condition has a unique solution $P(I)$. Given uniqueness of a solution and given the continuity of the market-clearing condition, it follows that $P(I)$ is continuous.

The set of v consistent with active intermediation. Let $\mathcal{V}(I)$ denote the function defined in equation (10). Note that, in any equilibrium, $P > 0$ implies by (16) that $q^\star(\underline{\delta}) > 0$ and thus $\mu I < s$. As a result, the set of v consistent with active intermediation is equal to the range of $\mathcal{V}(I)$ over the open interval $(0, s/\mu)$. As $I \to s/\mu$, customers' average asset holdings must go to zero, implying that the same is true for the smallest of customers' asset holdings, i.e., $q^\star(\underline{\delta} \mid P(I), I) \to 0$. Therefore, (16) implies that $u_q^{-1}(rP(I), \underline{\delta}) \to 0$ and, given Inada conditions, that $P(I) \to \infty$. It thus follows from (17) that $q^\star(\overline{\delta} \mid P(I), I) \to 0$ and that, for all (q', δ') in the support of the distribution $q' + I \geq q^\star(\underline{\delta} \mid P(I), I) + I > q^\star(\delta \mid P(I), I) \geq q^\star(\overline{\delta} \mid P(I), I)$. Hence, the inventory in advance never binds and $\mathcal{V}(I) = rP(I)$. We conclude that $\lim_{I \to s/\mu} \mathcal{V}(I) = +\infty$.

Next, recall that since the marginal trade surplus is decreasing in asset holdings and increasing in preference type, we have that

$$\Sigma(q', \delta') \leq \Sigma(q^\star(\underline{\delta} \mid P(I), I), \overline{\delta})$$

for all (q', δ') in the support of the stationary distribution. Keeping in mind that $\Sigma(q^\star(\underline{\delta} \mid P(I), I), \overline{\delta}) \geq 0$, we obtain from the Bellman equation of the marginal trade surplus that

$$\Sigma(q^\star(\underline{\delta} \mid P(I), I), \overline{\delta}) \leq \frac{u_q(q^\star(\underline{\delta} \mid P(I), I), \overline{\delta}) - rP(I)}{r + \gamma + \lambda(1 - \theta)} + \frac{\gamma + \lambda(1 - \theta)}{r + \gamma + \lambda(1 - \theta)} \Sigma(q^\star(\underline{\delta} \mid P(I), I), \overline{\delta}).$$

Simplifying and using the lower bound for asset holdings in equation (16), we obtain

$$\Sigma(q', \delta') \leq \frac{u_q\left(u_q^{-1}(rP(I), \underline{\delta}), \overline{\delta}\right) - rP(I)}{r}.$$

Now the upper bound on asset holdings (17), together with market clearing, implies $rP(I) \leq u_q(s - \mu I, \overline{\delta})$, since otherwise all asset holdings would be less than $s - \mu I$. Likewise, $rP(I) \geq u_q(s - \mu I, \underline{\delta})$. Taken together, we obtain an upper bound on the marginal trade surplus:

$$\Sigma(q', \delta') \leq \frac{u_q\left(u_q^{-1}(u_q(s - \mu I, \overline{\delta}), \underline{\delta}), \overline{\delta}\right) - u_q(s - \mu I, \underline{\delta})}{r}.$$

This implies that the marginal trade surplus remains bounded above as $I \to 0$ and that, in turn, $\mathcal{V}(I)$ remains bounded below as $I \to 0$. Altogether this means that

$$\underline{v} \equiv \inf_{I \in (0, s/\mu)} \mathcal{V}(I) > -\infty,$$

which completes the proof.

Note that, if we had instead assumed that dealers, like customers, had concave preferences with $\lim_{q \to 0} u'(q) = \infty$, then equilibria would always feature active intermediation. However, we think that risk neutral preferences are a more natural assumption for large dealers, and we view our model's implications for when (or under what conditions) markets feature active intermediation as an attractive feature of our framework.

B The Model Without Inventory Constraints

To better understand the unique implications of our model, we study the model without inventory-in-advance constraints. Note that Lemmas 2 and 3, below, are versions of existing results in Lagos and Rocheteau (2009) and Pinter and Üslü (2021). However, Lemmas 4 and 5—which establish results about the symmetry properties of this model—are new.

The results of this section are useful for several reasons. First, they provide natural initial conditions for our computations. Second, they help motivate the use of certain moments for parameter identification; in particular, they highlight the role played by the elasticity of the utility function in generating asymmetries in value-weighted transaction costs for purchases and sales. Third, they allow us to highlight unique properties of the model *with* inventory-in-advance constraint. In particular, in Lemma 4 we show that, without inventory constraints, the number of purchases and the number of sales are equal. Hence, in our model, any asymmetry in the number of sales and purchases is due to the presence of an inventory constraint.

B.1 Marginal surplus

In the model without inventory constraints, the Bellman equation for the marginal surplus reduces to

$$(r+\gamma+\lambda(1-\theta))\Sigma(q,\delta) = u_q(q,\delta) - rP + \gamma\mathbb{E}^F\left[\Sigma(q,\delta')\right].$$

Taking expectations on both sides we obtain that

$$\mathbb{E}^F\left[\Sigma(q,\delta')\right] = \frac{\mathbb{E}^F\left[u_q(q,\delta')\right] - rP}{r+\lambda(1-\theta)}.$$

Plugging back into the Bellman equation and simplifying yields the following results.

Lemma 2. *In the model without an inventory-in-advance constraint, the marginal surplus is*

$$(r+\lambda(1-\theta))\Sigma(q,\delta) = \frac{r+\lambda(1-\theta)}{r+\gamma+\lambda(1-\theta)}u_q(q,\delta) + \frac{\gamma}{r+\gamma+\lambda(1-\theta)}\mathbb{E}^F\left[u_q(q,\delta')\right] - rP.$$

In particular, for all $\eta > 0$, let $u(q,\delta) = \frac{q^{1-1/\eta}\delta}{1-1/\eta}$ if $\eta \neq 1$ and $u(q,\delta) = \log(q)\delta$ if $\eta = 1$. Then, the marginal surplus solves

$$(r+\lambda(1-\theta))\Sigma(q,\delta) = D(\delta)q^{-1/\eta} - rP,$$

where

$$D(\delta) \equiv \frac{r+\lambda(1-\theta)}{r+\gamma+\lambda(1-\theta)}\delta + \frac{\gamma}{r+\gamma+\lambda(1-\theta)}\mathbb{E}^F\left[\delta'\right].$$

B.2 Stationary distribution

In this Section, we derive a closed form solution for the steady-state distribution of asset holdings and preference shocks. Lagos and Rocheteau (2009) first proposed a solution for the case of a discrete distribution of preference shocks, and Pinter and Üslü (2021) extended this result to the case of an arbitrary distribution.

Lemma 3. *In the model without an inventory-in-advance constraint, the cumulative measure of customers with asset holdings less than q and utility type less than δ is*

$$\Phi(q',\delta') = \frac{\gamma}{\gamma+\lambda} F(\delta^\star(q'))F(\delta') + \frac{\lambda}{\gamma+\lambda} F(\min\{\delta^\star(q'),\delta'\}), \tag{19}$$

where $\delta^\star(q) \equiv (q^\star)^{-1}(\delta)$. In particular, for any integrable function $h(q',\delta')$, it holds that:

$$\begin{aligned}&\int_{(q',\delta')} h(q',\delta')\, d\Phi(q',\delta')\\ =&\frac{\gamma}{\gamma+\lambda}\int_{(\delta,\delta')} h(q^\star(\delta),\delta')\, dF(\delta)\, dF(\delta') + \frac{\lambda}{\gamma+\lambda}\int_{\delta'} h(q^\star(\delta'),\delta')\, dF(\delta').\end{aligned} \tag{20}$$

Perhaps the formula that is easiest to interpret is (20), which allows us to calculate any moment of the distribution. It shows that the joint distribution of asset holdings can be viewed as a mixture of two distributions. First, with probability $\gamma/(\gamma+\lambda)$, the distribution is random: a customer of type δ' is endowed with a random asset holding, $q^\star(\delta)$, where δ is drawn according to the distribution F. Second, with probability $\lambda/(\gamma+\lambda)$, the distribution is perfect: a customer of type δ' is endowed with her target asset holding, $q^\star(\delta')$.

Proof of equation (19). We characterize $\Phi(q,\delta)$ in two steps. First, we derive the measure of customers with holdings less than q,

$$\int_{(q',\delta')} \mathbb{I}_{\{q'\leq q\}}\, d\Phi(q',\delta').$$

In steady state, the gross outflow from the set of customers with holding less than q must equal the inflow:

$$\lambda\int_{(q',\delta')} \mathbb{I}_{\{q'\leq q\}}\, d\Phi(q',\delta') = \lambda\int_{\delta'} \mathbb{I}_{\{q^\star(\delta')\leq q\}}\, dF(\delta').$$

The left-hand side is the gross outflow, created by customers who contact the market with current holdings less than q. The right-hand side is the gross inflow, generated by all customers who contact dealers that have *optimal holdings* less than q. Clearly, $q^\star(\delta)\leq q$ if and only if $\delta\leq\delta^\star(q)$. Hence, the steady-state equation above can be written:

$$\int_{(q',\delta')} \mathbb{I}_{\{q'\leq q\}}\, d\Phi(q',\delta') = F(\delta^\star(q)).$$

This preliminary step facilitates the derivation of the entire distribution. Indeed, the inflow-outflow equation for $\Phi(q, \delta)$ can now be written:

$$(\gamma + \lambda)\Phi(q, \delta) = \gamma F(\delta^\star(q))F(\delta) + \lambda \int \mathbb{I}_{\{\delta' \le \delta \text{ and } q^\star(\delta) \le q\}} \, dF(\delta').$$

The left-hand side is the gross outflow, created by all customers with type less than δ and holdings less than q who either change type or contact dealers. The first term on the right-hand side is the gross inflow created by customers with holdings less than q who draw a new type less than δ. The second term is the gross inflow created by trade with dealers: customers with utility type less than δ and optimal holding $q^\star(\delta)$ less than q. Recalling the definition of $\delta^\star(q)$, equation (19) follows.

Proof of equation (20). By Theorem 3.6.1 in Bogachev (2007), it follows that:

$$\int_{q', \delta'} h(q', \delta') \, d\Phi(q', \delta') = \int_{\delta, \delta'} h(q^\star(\delta), \delta') \, d\Psi(\delta, \delta'),$$

where

$$\begin{aligned}
\Psi(\delta, \delta') &= \Phi(q^\star(\delta), \delta') \\
&= \frac{\gamma}{\gamma + \lambda} F(\delta)F(\delta') + \frac{\lambda}{\gamma + \lambda} F(\min\{\delta, \delta'\}) \\
&= \frac{\gamma}{\gamma + \lambda} \int_{\underline{\delta}}^{\delta} \int_{\underline{\delta}}^{\delta'} dF(x) \, dF(y) + \frac{\lambda}{\gamma + \lambda} \int_0^{\delta} dF(x) \mathbb{I}_{\{x \le \delta'\}} \\
&= \frac{\gamma}{\gamma + \lambda} \int_{\underline{\delta}}^{\delta} \int_{\underline{\delta}}^{\delta'} dF(x) \, dF(y) + \frac{\lambda}{\gamma + \lambda} \int_0^{\delta} dF(x) \int_0^{\delta'} d\mathbb{I}_{\{x \le y\}}.
\end{aligned}$$

The last line follows because $\mathbb{I}_{\{x \le \delta'\}} = \int_0^{\delta'} d\mathbb{I}_{\{x \le y\}}$, where $y \mapsto d\mathbb{I}_{\{x \le y\}}$ is the Dirac measure centered at point x. We thus obtain that:

$$d\Psi(\delta, \delta') = \frac{\gamma}{\gamma + \lambda} dF(\delta) \, dF(\delta') + \frac{\lambda}{\gamma + \lambda} dF(\delta) d\mathbb{I}_{\{\delta \le \delta'\}},$$

and equation (20) follows.

B.3 Buy-sell symmetry

In this section, we show that, in the model without inventory constraints, the number of purchases and sales are equal. With a continuum of customers, the natural measure of "number of purchases" is the flow of purchases per unit of time:

$$\lambda \int_{(q',\delta')} \mathbb{I}_{\{q'<q^\star(\delta')\}}\, d\Phi(q',\delta').$$

Using equation (20) with $h(q',\delta') = \mathbb{I}_{\{q^\star(\delta')>q'\}}$, we obtain that the flow of purchase is

$$\frac{\lambda\gamma}{\gamma+\lambda}\int_{(\delta,\delta')} \mathbb{I}_{\{q^\star(\delta)<q^\star(\delta')\}}\, dF(\delta)\, dF(\delta') + \frac{\lambda^2}{\gamma+\lambda}\int_{\delta'} \mathbb{I}_{\{q^\star(\delta')<q^\star(\delta')\}}\, dF(\delta').$$

Since the target holding function is strictly increasing, the indicator in the first integral simplifies to $\mathbb{I}_{\{\delta<\delta'\}}$. Moreover, it is clear that the indicator in the second integral is zero. Hence, the flow of purchase is

$$\frac{\lambda\gamma}{\lambda+\gamma}\int_{(\delta,\delta')} \mathbb{I}_{\{\delta<\delta'\}} dF(\delta)\, dF(\delta') = \frac{\lambda\gamma}{\lambda+\gamma}\int_{\delta} dF(\delta)\,(1-F(\delta)) = \frac{\lambda\gamma}{2\,(\lambda+\gamma)}\left(1-\sum_{\delta\in[\underline{\delta},\overline{\delta}]}\Delta F(\delta)^2\right),$$

where $\Delta F(\delta) = F(\delta) - F(\delta-)$ represents the jump (if any) in the CDF $F(\cdot)$ at δ and the last equality follows from the integration by parts formula for functions of bounded variations, which can be found in Theorem 6.2.2 of Carter and Van Brunt (2000). Following the same steps reveals that the flow of sales is

$$\lambda \int_{(q',\delta')} \mathbb{I}_{\{q'>q^\star(\delta')\}}\, d\Phi(q',\delta') = \frac{\lambda\gamma}{\lambda+\gamma}\int_{(\delta,\delta')} \mathbb{I}_{\{\delta>\delta'\}}\, dF(\delta)\, dF(\delta'),$$

which is clearly equal to the flow of purchases. The following lemma summarizes.

Lemma 4. *In the model without an inventory-in-advance constraint, the flow of purchases and the flow of sales are both equal to*

$$\frac{\lambda\gamma}{2(\lambda+\gamma)}\left(1-\sum_{\delta\in[\underline{\delta},\overline{\delta}]}\Delta F(\delta)^2\right).$$

Correspondingly, the average trade size of a sale and of a purchase are also equal.

B.4 Transaction cost asymmetry

We now investigate a different source of asymmetry: the average proportional transaction cost incurred by customers who purchase assets vs. those who sell it. We show that, in the model without an inventory-in-advance constraint, the transaction costs are in general asymmetric. In addition, the direction of the asymmetry depends on the elasticity of the utility function $u(q,\delta)$.

Let W denote the total value of purchases which, by market clearing, must be equal to the total value of sales. The value-weighted proportional transaction costs for purchases can be written:

$$
\begin{aligned}
\mathrm{TC}_p =& \int_{(q',\delta')} \mathbb{I}_{\{q^\star(\delta')>q'\}} \frac{P\left(q^\star(\delta')-q'\right)}{W} \theta \frac{V(q^\star(\delta'),\delta') - V(q',\delta') - P(q^\star(\delta')-q')}{q^\star(\delta')-q'} \, d\Phi(q',\delta') \\
=& \frac{\gamma}{\gamma+\lambda} \frac{\theta P}{W} \int_{\delta<\delta'} \left[V(q^\star(\delta'),\delta') - V(q^\star(\delta),\delta') - P(q^\star(\delta')-q^\star(\delta))\right] \, dF(\delta) dF(\delta'),
\end{aligned}
$$

where the second equality follows from equation (20). This formula shows that the value-weighted transaction cost is proportional to the total surplus generated by purchases. Likewise, we obtain that the value-weighted proportional transaction costs for sales writes:

$$
\begin{aligned}
\mathrm{TC}_s =& \int_{(q',\delta')} \mathbb{I}_{\{q^\star(\delta')<q'\}} \frac{P\left(q'-q^\star(\delta')\right)}{W} \theta \frac{V(q^\star(\delta'),\delta') - V(q',\delta') - P(q^\star(\delta')-q')}{q'-q^\star(\delta')} \, d\Phi(q',\delta') \\
=& \frac{\gamma}{\lambda+\gamma} \frac{\theta P}{W} \int_{\delta'<\delta} \left[V(q^\star(\delta'),\delta') - V(q^\star(\delta),\delta') - P(q^\star(\delta')-q^\star(\delta))\right] \, dF(\delta) dF(\delta') \\
=& \frac{\gamma}{\lambda+\gamma} \frac{\theta P}{W} \int_{\delta<\delta'} \left[V(q^\star(\delta),\delta) - V(q^\star(\delta'),\delta) - P(q^\star(\delta)-q^\star(\delta'))\right] \, dF(\delta) dF(\delta') \\
=& -\frac{\gamma}{\lambda+\gamma} \frac{\theta P}{W} \int_{\delta<\delta'} \left[V(q^\star(\delta'),\delta) - V(q^\star(\delta),\delta) - P(q^\star(\delta')-q^\star(\delta))\right] \, dF(\delta) dF(\delta'),
\end{aligned}
$$

where the second to last line renames the variables of integration, replacing δ with δ' and vice versa.

Subtracting the above expressions for TC_p and TC_s, and writing the surplus as an integral of the marginal surplus, we see that $\mathrm{TC}_p > \mathrm{TC}_s$ if, for all $\delta < \delta'$,

$$
\int_{q^\star(\delta)}^{q^\star(\delta')} \left[\Sigma(q,\delta) + \Sigma(q,\delta')\right] \, dq > 0. \tag{21}
$$

Alternatively, $\mathrm{TC}_p < \mathrm{TC}_s$ if the above expression is strictly negative for all $\delta < \delta'$.

For intuition, consider any pair of target holdings $q^\star(\delta)$ and $q^\star(\delta')$, with $\delta < \delta'$. A customer will purchase the quantity $q^\star(\delta') - q^\star(\delta)$ if her current asset holding is $q^\star(\delta)$ and her current utility type is δ'. With Nash-bargaining, transaction costs are proportional to the surplus, which can be calculated by integrating below the marginal surplus curve, $\Sigma(q,\delta')$. Likewise, a customer will sell the same quantity $q^\star(\delta') - q^\star(\delta)$ when her current asset holding is $q^\star(\delta')$ and her current utility type is δ. In that case, the transaction cost is obtained by integrating below the marginal surplus curve $-\Sigma(q,\delta)$. Condition (21) ensures that the integral is larger for the purchase than for the sale.

Now, we explore the impact of parameter values in determining whether total transaction costs will be larger for purchases or sales. We focus on an isoelastic utility function of the form, $u(q,\delta) = \frac{q^{1-1/\eta}}{1-1/\eta}\delta$, identical to the one we use in the quantitative exercise.

Lemma 5. *Suppose that the utility function is isoelastic, $u(q,\delta) = \frac{q^{1-1/\eta}}{1-1/\eta}\delta$. Then:*

$$\mathrm{TC}_p \begin{cases} > \mathrm{TC}_s & \text{if } \eta < 2 \\ = \mathrm{TC}_s & \text{if } \eta = 2 \\ < \mathrm{TC}_s & \text{if } \eta > 2. \end{cases}$$

With an isoelastic utility function, the marginal surplus has a simple closed-form solution shown in Lemma 2. Then, condition (21) writes:

$$\frac{1}{2}\left(D(\delta) + D(\delta')\right)\left(\frac{q^\star(\delta')^{1-1/\eta}}{1-1/\eta} - \frac{q^\star(\delta)^{1-1/\eta}}{1-1/\eta}\right) - rP\left(q^\star(\delta') - q^\star(\delta)\right) > 0.$$

Since target holdings have zero marginal surplus, we have that $D(\delta)q^\star(\delta)^{-1/\eta} = D(\delta')q^\star(\delta')^{-1/\eta} = rP$, implying that target holdings are

$$q^\star(\delta) = \left(\frac{D(\delta)}{rP}\right)^{-\eta} \text{ and } q^\star(\delta') = \left(\frac{D(\delta')}{rP}\right)^{-\eta}.$$

Plugging back, and assuming for now that $\eta \neq 1$, we can factor out the price rP and, after letting $x \equiv D(\delta')/D(\delta)$, we find that condition (21) holds for all $\delta' > \delta$ if and only if

$$f(x) > 0 \text{ for all } x > 1, \text{ where } f(x) \equiv \frac{x+1}{2}\frac{x^{\eta-1}-1}{1-1/\eta} - \left(x^\eta - 1\right).$$

Taking derivatives twice, we obtain that:

$$\begin{aligned} \frac{df}{dx} &= \frac{\eta}{2}\left(\frac{x^{\eta-1}-1}{\eta-1} + x^{\eta-2} - x^{\eta-1}\right) \\ \frac{d^2 f}{dx^2} &= \frac{\eta}{2}x^{\eta-3}\left(2-\eta\right)(x-1). \end{aligned}$$

Hence, $df/dx(x) = 0$ when $x = 1$, is strictly increasing in x if $\eta < 2$, is identically equal to zero if $\eta = 2$, and is strictly decreasing if $\eta > 2$. Since $f(1) = 0$ as well, it thus follows that, for $x > 1$, $f(x) > 0$ if $\eta < 2$, $f(x) = 0$ if $\eta = 2$ and $f(x) < 0$ if $\eta > 2$. The result follows. The log case $\eta = 1$ can be addressed separately.

C Sell-to-Buy Ratio: Adjustment for Order Imbalance

Suppose that, over some period of time, we index customer purchases by $b \in \{1, 2, ..., N_B\}$ and customer sales by $s \in \{1, 2, ..., N_S\}$. The total quantity of customer purchases and sales are then

$$Q_B = \sum_{b=1}^{N_B} q_b, \quad \text{and} \quad Q_S = \sum_{s=1}^{N_S} q_s,$$

where q is the trade quantity. The average trade size for purchases and sales are

$$\overline{Q}_B = \frac{1}{N_B} \sum_{b=1}^{N_B} q_b, \quad \text{and} \quad \overline{Q}_S = \frac{1}{N_S} \sum_{s=1}^{N_S} q_s.$$

If the market is "balanced" during that time period—i.e., if $Q_B = Q_S$ and dealers do not accumulate net inventories—then we have

$$N_B \overline{Q}_B = N_S \overline{Q}_S \implies \frac{N_S}{N_B} = \frac{\overline{Q}_B}{\overline{Q}_S}.$$

However, in the data, the market is not always balanced. To adjust for order imbalance, let us define

$$\hat{N}_S = N_S \times \min\left\{\frac{Q_B}{Q_S}, 1\right\}, \quad \text{and} \quad \hat{N}_B = N_B \times \min\left\{\frac{Q_S}{Q_B}, 1\right\}.$$

Suppose, for example, that $Q_S > Q_B$, so that the quantity Q_B is sold to customers and the remaining quantity $Q_S - Q_B > 0$ is absorbed onto dealers' balance sheets to clear the market. If we further assume that the average trade size is the same for both portions of the sale volume ($Q_S - Q_B$ and Q_B), then $\hat{N}_S$ and $\hat{N}_B$ represent the number of customer sales and purchases that do *not* stay on dealers' balance sheets.

We can see that, by construction,

$$\frac{\hat{N}_S}{\hat{N}_B} = \frac{\overline{Q}_B}{\overline{Q}_S}.$$

In Figure 6, we plot the adjusted ratio $\frac{\hat{N}_S}{\hat{N}_B}$.