Over-the-Counter Intermediation, Customers' Choice and Liquidity Measurement*

Mariano J. Palleja †

October 29, 2023.

[Click here for the latest version]

Abstract

Stringent financial regulations and emerging technologies have reshaped over-the-counter intermediation, discouraging dealers from providing immediacy to customers using their own inventories (principal trading) in favor of a larger matchmaking activity (agency trades). This paper studies how customers optimally choose between these trading mechanisms and the implications of this choice over market liquidity. I develop a quantitative search model where heterogeneous customers choose between immediate but expensive and delayed but less costly trades, i.e., principal and agency trades, respectively. Each customer solves this speed-cost trade-off, jointly determining her optimal mechanism, transaction costs, and trading volume. When market conditions change, customers migrate across mechanisms in pursuit of higher trading surpluses. I show that this migration is not random, thus liquidity measures change not only because the market conditions did, but also because of a composition effect. To quantify such an effect, I structurally estimate my model and build counterfactual measures that control for migration. I replicate the major innovations seen in these markets and find that composition effects explain more than a third of the increase in principal transaction costs.

*I am very grateful to my advisors Pierre-Olivier Weill, Saki Bigio, Andy Atkeson, and Lee Ohanian for their invaluable guidance, advice, encouragement, and support throughout this research project. I also thank Mahyar Kargar and David Baqaee for their useful suggestions and advice. I thank Yesol Huh, Michael Gordy, Valery Polkovnichenko, and Andreas Rapp for their mentorship during my Dissertation Fellowship at the Federal Reserve Board. The paper has also benefited from long discussions with Fatih Ozturk and Luis Cabezas, to whom I owe a huge amount of gratitude. I thank Yang-Ho Park, Erfan Danesh, Xin Huang, Pawel Szerszen, Daniel Covitz, Patrick McCabe, Borghan Narajabad, Sebastian Infante, Brianna Chang, Marius Zoican and all seminar participants at UCLA, the Federal Reserve Board, the 2023 Econometric Society European Meeting and the 2023 Northern Finance Association Conference for their useful comments.

†UCLA, Department of Economics. Contact: marianopalleja@g.ucla.edu, Website: www.marianopalleja.com.

1 Introduction

Over-the-counter (OTC) markets are characterized by the lack of a centralized exchange in which customers can trade securities. Instead, customers need to search for trading counterparties. Dealers mitigate these search frictions in two ways. First, by trading with customers using their own inventories, i.e., by performing principal trades. Second, by matching customers with offsetting liquidity needs, i.e., by performing agency trades[1]. These two trading mechanisms, principal and agency, represent for customers a speed-cost trade-off. Principal trades are immediate but, given the implied inventory costs, are also costly. Conversely, agency trades are cheaper but imply an execution delay, caused by the time it takes to find a suitable counterparty.

Post-2008 financial regulations and recent technological changes have had a major impact on the relative cost of supplying these two types of trades. The implementation of the Dodd-Frank Act and the Basel III framework increased dealers' inventory costs, reducing their willingness to trade on a principal basis (Duffie, 2012; Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018). Quoting Goldman Sachs: *"Banks are committing less capital to trading desks with fixed income assets down 22% since 2010, and have exited some businesses altogether; for example, J.P. Morgan and Morgan Stanley no longer make markets in physical commodities while Deutsche Bank has exited single-name CDS"*[2]. In turn, the rising popularity of electronic trading venues made matching customers easier, shifting intermediation further away from dealers' inventories (O'Hara and Zhou, 2021).

Although the literature has extensively studied the dealers' optimal intermediation strategy when market conditions change, the customers' optimal response to such strategy and its implications for liquidity measurement have remained relatively unexplored. Notably, the speed-cost trade-off previously described suggests that customers may optimally migrate across trading mechanisms when market conditions change. Moreover, the decentralized nature of OTC markets – in which each customer bargains her own terms of trade – suggests that this migration might affect liquidity measures, by altering the samples over which these measures are computed.

In this paper, I develop and estimate a quantitative search model where I explicitly study the trading mechanism choice of each customer. I use this model to address how this trading mechanism choice affects liquidity measures when market conditions change. The model features risk-averse customers choosing between immediate but expensive and delayed but less costly trades, i.e., principal and agency trades, respectively. I find that customers with larger trading needs choose to buy and sell on principal. Intuitively, when trading is relatively urgent, the immediacy benefit outweighs the principal premium paid. Furthermore, customers with larger trading needs pay higher transaction costs, given that dealers extract higher fees from them. When market conditions change a fraction of customers optimally migrate across trading mechanisms.

[1] Agency trades are also known in the literature as riskless principal or matchmaking trades. The key characteristic of this mechanism is that the dealer avoids involving her own inventories by pre-arranging both legs before executing them.

[2] Goldman Sachs Global Investment Research, August 2, 2015 Report.

Therefore, principal and agency transaction cost measures change not only because the market conditions did, but also because of a composition effect. To quantify this composition effect, I develop counterfactual measures of transaction costs that control for migration. I structurally estimate the model using corporate bond transaction data and revisit the two major innovations this market suffered in the last decade. I find that the standard practice of comparing average transaction costs before and after a market condition change overestimates the impact of these changes. Specifically, composition effects account for 32% of the rise in principal costs after an inventory costs increase and for around 90% of the change after an increase in the agency execution speed. In turn, agency costs are barely affected by composition effects.

The model developed explicitly accounts for the optimal decisions of customers facing alternative trading mechanisms in OTC markets. Particularly, I build on the framework in Lagos and Rocheteau (2009) (hereafter LR09). The model features search frictions, heterogeneous risk-averse customers trading a perfectly divisible asset, and bilateral bargaining over the terms of trade. My theoretical contribution relative to LR09 is that I allow customers to choose between two trading mechanisms, which resemble principal and agency trades in practice. Principal trading is immediate but costly. This responds to dealers partially translating their implied inventory costs to customers. Agency trading is delayed but cheaper: finding a suitable counterparty takes time, but dealers avoid incurring inventory costs. These features enable me to study the speed-cost trade-off aforementioned.

I find that, in equilibrium, customers sort themselves across mechanisms depending on their liquidity needs. Customers with a larger distance between current and optimal asset positions choose to trade on principal. Conversely, customers with positions closer to their optimal ones choose to wait for an agency execution. The explanation relies on customers obtaining a marginally decreasing utility from holding assets. The bigger the distance between customers' current and optimal positions, the higher their marginal trading surplus and the higher their willingness to pay for an immediate execution.

This optimal sorting has a direct impact on liquidity measures. In the model, optimal mechanisms and transaction costs are jointly determined. Specifically, transaction costs are bargained, and thus they incorporate a customer's specific trading surplus. The more a customer needs to trade, the larger the marginal trading surplus she attains and the higher the cost she has to pay for each unit traded. As can be seen, when trading needs are large, not only are customers more likely to opt for the principal trade, but they also pay higher transaction costs. The implication is that principal traders pay on average higher costs not only because of the inventory costs implied by such a mechanism but also because of selection: customers trading on principal have on average larger trading needs than those trading on agency.

I use this framework to analyze the optimal reaction of customers when market conditions change and its implications for liquidity measurement. Specifically, I consider changes in the two key parameters that affect the speed-cost trade-off faced by customers: the inventory costs implied by principal trades and

the execution speed of agency trades. These changes resemble recent market innovations, where stricter regulations increased inventory costs and the rising popularity of electronic trading venues eased agency trading. Not surprisingly, in both cases, customers endogenously migrate away from principal trading. Furthermore, such migration is not random: among principal traders, only those with smaller trading needs migrate towards agency. Intuitively, smaller trading needs place customers closer to being indifferent between principal and agency trading, given that the marginal surplus from fast trading is closer to the premium cost paid for it.

Such a heterogeneous response implies an empirical issue when trying to estimate the impact of a market innovation on liquidity. In this regard, the empirical literature has widely exploited the relation between trading mechanisms and execution delays to overcome a recurrent inconvenience: execution delays are not observed. Particularly, when measuring transaction costs, researchers would split trades beforehand according to the trading mechanism used. Principal costs would account for the price of immediacy, whereas agency costs would measure the price of delayed executions [3]. Although splitting trades in such a way purges transaction cost measures from execution delay changes, it overlooks the fact that the obtained samples are endogenous: they are the result of a choice. When market conditions change, customers endogenously migrate, and thus the estimates of the impact on a mechanism's transaction costs are subject to a composition bias. For example, an increase in inventory costs would reduce the sample of principal traders to those with higher trading needs. In such a case, the effect of increasing inventory costs on principal transaction costs would be overestimated. This bias can hardly be narrowed when the characteristics in which the samples differ cannot be observed.

Equipped with the steady-state equilibrium of my model, I tackle this empirical issue. Firstly, I decompose the equilibrium distribution of customers into those that, after a market innovation, continue using the same mechanism or not, i.e., the non-migrant and migrant customers, respectively. Secondly, for each mechanism I compute measures of transaction cost changes, using both the entire distribution of customers before and after the innovation, as well as the subset of non-migrant customers. The comparison of these measures returns the sign and size of the composition bias.

To ensure that my numerical results are grounded in the data, I structurally estimate the model. For this, I use transaction data on the US corporate bond secondary market. Specifically, I employ the academic version of the Trade Reporting and Compliance Engine (TRACE) database from January 2016 to December 2019. Importantly, this data contains dealers' identifiers, thus it allows me to distinguish between principal and agency trades. I target a set of relevant empirical moments and use the generalized method of moments

[3] There are two main strategies to identify principal and agency trades. The first one infers agency trades as those offsetting transactions performed by the same dealer within a small time window (usually between one and fifteen minutes), labeling as principal all remaining trades (Schultz, 2017; Goldstein and Hotchkiss, 2020; O'Hara and Zhou, 2021; Choi, Huh, and Shin, 2023). A second method is to isolate episodes where arguably only principal trades are performed, such as downgrades (Bao, O'Hara, and Zhou, 2018), extreme market volatility events (Anderson and Stulz, 2017), or index exclusions (Dick-Nielsen and Rossi, 2019).

to jointly estimate the deep parameters of the model.

Finally, the estimated model is used to revisit the empirical evidence related to the transaction costs evolution after two major OTC markets' innovations. I perform numerical exercises that replicate both the introduction of post-2008 stricter financial regulations and the rise of electronic trading venues. In both cases, when the economic environment changes, migration across mechanisms happens. Using the aforementioned strategy, I show that the composition bias matters: it explains an economically significant fraction of the change in transaction costs.

Regarding the first exercise proposed, the aftermath of the 2008 financial crisis saw the introduction of new regulations aimed at increasing the financial market's resilience. The adoption of the Dodd-Frank Act in the United States and the Basel III framework internationally, restrictions meant to reduce banks' exposure to risky assets, negatively affected their dealership activity. Specifically, these new regulations increased banks' cost of holding assets in their balance sheets, thus reducing their willingness to provide liquidity on a principal basis (Duffie, 2012). Several papers have addressed the impact of these new regulations on market transaction costs. Overall, the consensus is that principal costs have increased since the new regulations took place, with intermediation shifting away from principal trading towards larger agency activity (Anderson and Stulz, 2017; Schultz, 2017; Bao, O'Hara, and Zhou, 2018; Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; Dick-Nielsen and Rossi, 2019; Choi, Huh, and Shin, 2023). I analyze such an increase in inventory costs through the lens of the model. The exercise suggests that previous estimates overstate the increase in principal costs. Particularly, I find that the composition bias accounts for a third of the increase in principal costs while it does not play an economically significant role in the change of agency costs.

The second numerical exercise is motivated by the emergence of electronic trading venues. In contrast with traditional voice trading, electronic requests for quotes allow customers to contact multiple dealers at the same time. The empirical evidence tells us that the agency share is higher for those bonds that are traded electronically and that dealers use electronic platforms to find counterparties for customers that contacted them through traditional voice messages (Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; O'Hara and Zhou, 2021). From the customers' perspective, the rising popularity of electronic trading venues implies that dealers can match them with a counterparty faster. To replicate this market innovation, I reduce the expected agency execution delay of the model. I find that transaction costs increase in both mechanisms. However, while the composition bias implies a negligible underestimation of the change in agency costs, it explains most of the increase in principal transaction costs.

Overall, the results in this paper suggest that taking into account customers' optimal response can better inform policymakers about the impact that market innovations have on market liquidity. Firstly, this is because customers optimally migrate across mechanisms, mitigating the effect of worsening conditions and fostering the effects of improving ones. Secondly, considering the customers' response allows us to better

measure the impact of these new conditions. In particular, I show that stricter financial regulations have not increased principal transaction costs as much as was previously thought.

1.1 Related Literature

This paper develops a theoretical model of trading mechanism choice in OTC markets that allows me to revisit quantitatively recent evidence on transaction cost changes. It contributes to three strands of the literature.

Firstly, this paper contributes to the search literature in OTC markets, pioneered by Duffie, Gârleanu, and Pedersen (2005) and Lagos and Rocheteau (2009), and summarized in Weill (2020). In this literature, when customers and dealers meet, execution is immediate. I relax this assumption by explicitly modeling two trading mechanisms, which resemble principal and agency trades in practice. This feature allows me to study theoretically the customers' trade-off between expensive but immediate and cheaper but slower execution. I show that the optimal mechanism choice can be characterized by preference-specific asset holdings thresholds, and analyze how such thresholds change according to the key parameters of the model. In their independent, contemporaneous work, Dyskant, Silva, and Sultanum (2023) also include alternative trading mechanisms in a search model. In their framework, customers are restricted to holding either zero or one unit of the asset. In contrast, I allow for unrestricted asset holdings and show that the endogenous trade size of each customer determines her trading mechanism choice. I further exploit the relation between trade size and transaction costs to estimate my model and perform quantitative exercises where I assess the role that migration plays when measuring liquidity.

This paper also contributes to the theoretical literature that explicitly accounts for principal and agency trading in OTC markets (Cimon and Garriott, 2019; Plante, 2021; An, 2022; An and Zheng, 2023; Saar, Sun, Yang, and Zhu, 2023). This literature addresses how dealers manage their inventories by setting the optimal principal trade cost: if the principal cost increases customers migrate towards agency trading, reducing the inventory burden [4]. In my model, both the trading mechanism choice and the terms of trade in each mechanism are the results of bilateral bargaining between dealers and customers. The consequences are twofold. First, it provides a non-degenerate distribution of transaction costs within each trading mechanism, which I exploit to estimate the model. This is because the terms of trade reflect both the incurred cost of the bargaining dealer and the trading surplus of the bargaining customer. Second, it allows me to study how composition effects affect liquidity measures in a quantitative way. In line with the existing literature, when the principal premium increases the sample of customers trading on principal reduces. In contrast with the existing literature, the reduction of the sample does affect the average principal transaction costs, given that each customer bargains her own transaction cost.

[4] A less related literature studies the customers' optimal choice of trading in a centralized or a decentralized market (Miao, 2006; Shen, 2015)

Finally, this paper complements the empirical literature that addresses transaction cost changes and trading mechanism shifts in OTC markets. It has been documented that the regulation set after the 2008 financial crisis changed the liquidity profile of the corporate bond market. Specifically, researchers have shown that principal trading is less abundant and more costly (Anderson and Stulz, 2017; Schultz, 2017; Bao, O'Hara, and Zhou, 2018; Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; Dick-Nielsen and Rossi, 2019; Choi, Huh, and Shin, 2023; Rapp and Waibel, 2023). Additionally, the empirical evidence indicates that the rising popularity of electronic trading venues had attracted volume towards agency trading, reducing the cost of such trades (Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; O'Hara and Zhou, 2021). Finally, during episodes of big turmoil, e.g., COVID-19, researchers have documented a rise in the cost of principal trading with an associated shift away from it (Kargar, Lester, Lindsay, Liu, Weill, and Zúñiga, 2021). A common feature across these papers is the lack of customer data, which prevents them from controlling the documented customers' endogenous migration when computing transaction cost changes [5]. I complement these papers by analyzing the sign and size of the consequent composition bias. To achieve this goal, I exploit the model to construct counterfactual distributions in which transaction cost changes can be measured using a steady sample of customers. I show that the estimates of transaction cost changes provided by this literature include an economically significant composition bias, and thus can hide the true speed-cost trade-off customers face.

2 The Model

In this section I explain the model. I start by describing the environment and the problems that both customers and dealers face. Later I show how terms or trade are set, highlighting the link between transaction costs and trading mechanism choice. Finally, I define the steady-state equilibrium.

2.1 Environment

I build on LR09 continuous time model of an OTC secondary market with search frictions. There is a single asset in fixed supply $A \in \mathbb{R}_+$, and two types of infinitely lived agents: customers and dealers, both in unit measure and discounting time at rate $r > 0$. Customers hold an asset in quantity $a \in \mathbb{R}_+$ and derive utility from two different consumption goods, *fruit* and *numéraire*. *Fruit* is perishable, non-tradable, and produced by the asset in a one-to-one ratio. In turn, the *numéraire* good is produced by all agents. The instantaneous utility function of a customer is $u_i(a) + d$, where a and d represent the consumption of *fruit* and the net consumption of the *numéraire* good, respectively, and $i \in \{1, ..., I\}$ indexes the preference type.

[5] Goldstein and Hotchkiss (2020) study corporate bonds' inventory risk, and address the endogeneity of trading mechanisms by implementing an endogenous switching regression. Given that their data does not contain customer characteristics, they use bond and trade characteristics to predict the optimal mechanism of a trade.

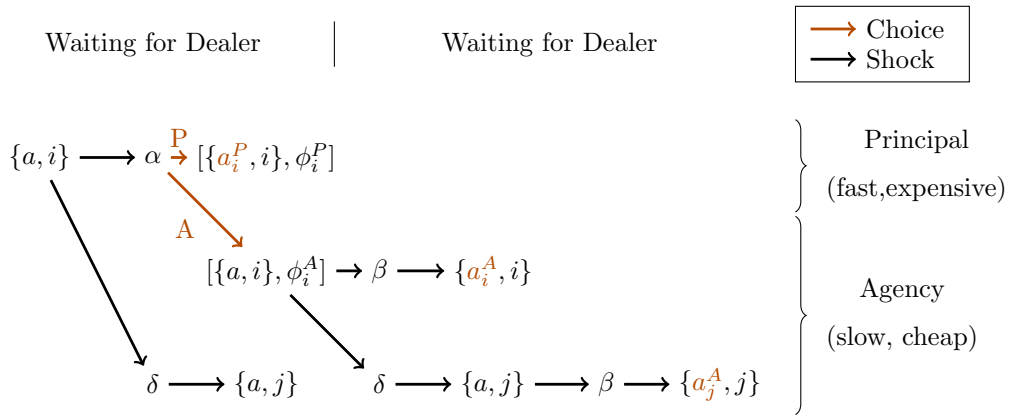
Specifically, the instantaneous utility provided by *fruit* is assumed iso-elastic, $u_i(a) = \epsilon_i \times a^{1-\sigma}/(1-\sigma)$, with multiplicative preference shifters ϵ_i. Each customer is subject to an independent preference shock process, which follows a Poisson distribution with arrival rate δ. Once hit by the preference shock, a new type i is assigned with probability π_i, where $\sum_{i=1}^{I} \pi_i = 1$. This change in preferences creates a motive for trade in the model, and can be interpreted as changing hedging needs (Duffie, Gârleanu, and Pedersen, 2007; Vayanos and Weill, 2008), changing beliefs about the asset's future payoff (Hugonnier, 2012), etc.

Customers can trade assets only when they contact a dealer, an event that is governed by a Poisson process with an arrival rate of α. Once a customer meets a dealer, she chooses among two kinds of trading mechanisms: principal or agency, denoted by superscripts P and A, respectively. On the one hand, if she opts for the principal trade, she immediately exchanges each unit of her excess position at the inter-dealer price p and pays a transaction cost of ϕ^P. On the other hand, if she opts for an agency trade, she waits until the dealer finds her a counterparty, and meanwhile enjoys the utility provided by her current asset holdings. It is assumed that she will be matched at a random time according to a Poisson process with β arrival rate. When matched, this customer rebalances her position at p and pays the dealer a transaction cost ϕ^A. I further assume that a customer cannot contact any other dealer while she is waiting for her trade to be executed. Thus, at every moment, customers will be either waiting to contact a dealer or waiting for their agency trade to be executed. These two states are denoted by ω_1 and ω_2, respectively.

Transaction costs and quantities are determined through a Nash bargaining protocol that takes place at the moment of contact with the dealer. This timing assumption implies that, for agency trades, the negotiation is based on the expected trade surplus a customer subject to preferences shocks might achieve. More details about these terms of trade are presented in subsection 2.2. After transactions are completed, the dealer and the customer part ways.

At any time, customers find themselves with certain asset holdings a_t, preference type i_t, and within a specific waiting state ω_t. Thus, customers can be fully characterized by the triplet $\{a_t, i_t, \omega_t\} \in \mathcal{O}$, where $\mathcal{O} = \mathbb{R}_+ \times \{1, ..., I\} \times \{\omega_1, \omega_2\}$. This heterogeneity is depicted with a probability space $(\mathcal{O}, \Sigma, H_t)$, where Σ is the σ-field generated by the sets $(\mathcal{A}, \mathcal{I}, \mathcal{W})$, with $\mathcal{A} \subseteq \mathbb{R}_+$, $\mathcal{I} \subseteq \{1, ..., I\}$, $\mathcal{W} \subseteq \{\omega_1, \omega_2\}$, and H_t is a probability measure on Σ that represents the distribution of customers across the state space at time t. Figure 1 outlines a customer's potential paths from the moment she contacts a dealer until she executes her trade.

Figure 1: Customer Path.



Note: This figure shows a customer's path through the state space. Shocks are depicted by black arrows, and include the contact with dealers (α), the change of preference (δ), and the execution of the agency trade (β). The customer's choice is depicted in orange arrows and includes the optimal trading mechanism and the corresponding new asset holdings.

Since I am going to focus on the steady-state equilibrium, to simplify the notation I disregard the time dependence when it is not strictly necessary. The maximum expected discounted utility attainable by a customer waiting for a dealer with preference type i at time t and asset holding a, $V_{i(t)}(a)$, satisfies

$$V_{i(t)}(a) = \mathbb{E}_{i(t)}\Big[\int_t^{T_\alpha} e^{-r(s-t)}u_{i(s)}(a)ds + e^{-r(T_\alpha - t)}\max\Big\{V^P_{i(T_\alpha)}(a), V^A_{i(T_\alpha)}(a)\Big\}\Big], \tag{1}$$

where

$$V^P_{i(T_\alpha)}(a) = V_{i(T_\alpha)}(a^P_{i(T_\alpha)}) - p(a^P_{i(T_\alpha)} - a) - \phi^P_{i(T_\alpha)}(a),$$

$$V^A_{i(T_\alpha)}(a) = \int_{T_\alpha}^{T_\beta} e^{-r(s-T_\alpha)}u_{i(s)}(a)ds + e^{-r(T_\beta - T_\alpha)}\Big[V_{i(T_\beta)}(a^A_{i(T_\beta)}) - p(a^A_{i(T_\beta)} - a) - \phi^A_{i(T_\alpha)}(a)\Big].$$

T_α and T_β are the next time a customer contacts a dealer and the execution time of the agency trade, respectively. The expectation operator $\mathbb{E}_{i(t)}$ is over the arrival times of contact with dealers, the execution of the agency trade, and the expected stream of preference types $i(s)$, conditional on the customer being of a certain preference type at t. Transaction costs and prices are expressed in units of the *numéraire* good.

Note that the optimal asset holdings under the two trading mechanisms, $a^P_{i(T_\alpha)}$ and $a^A_{i(T_\beta)}$, might differ for two reasons. Firstly, a customer might change her type during the waiting period of a delayed trade. Hence, types $i(T_\alpha)$ and $i(T_\beta)$ might be different. Secondly, the transaction costs charged by dealers in each kind of trade might differ independently of the aforementioned reason: each trading mechanism will require the dealer to face a different cost. Since transaction costs add up to the effective price of a trade, customers may choose different optimal asset holdings in different mechanisms.

In turn, dealers trade on behalf of their customers in the inter-dealer market. If they are asked to execute a principal trade, they need to incur a cost $\theta \in [0, \frac{r}{r+\beta})$ per (*numeraire*) dollar traded. In line with existing literature (e.g., An and Zheng, 2023; Saar, Sun, Yang, and Zhu, 2023), I assume that dealers' marginal inventory costs are constant. In this regard, Duffie et al. (2023) shows that liquidity measures are not affected by the level of dealers' inventory capacity utilization unless the latter is at an abnormally high level. Thus, the assumption is empirically supported as such a scenario of extremely high capacity utilization is not considered [6]. On the other hand, if the client asks the dealer to perform an agency trade, they wait until a counterparty is found, and the transaction cost is charged at execution. A representative dealer does not hold positions and her instantaneous utility equals her consumption of the *numéraire* good. Thus, her expected utility is given by the present value of the transaction costs she collects net of the costs she incurs. A dealer's maximum expected discounted utility satisfies

$$W(t) = \mathbb{E}\Big[e^{-r[T_\alpha - t]}\Big(\int_{\mathcal{O}} \Phi_{i(T_\alpha)}(a) dH_{T_\alpha} + W(T_\alpha)\Big)\Big], \tag{2}$$

where $\Phi_i(a) = \mathbf{1}_{[\text{P trade}]}\Big(\phi_i^P(a) - \theta p|a_i^P - a|\Big) + \mathbf{1}_{[\text{A trade}]}\Big(e^{-(T_\beta - T_\alpha)}\phi_i^A(a)\Big)$ and the integration over the probability measure H_{T_α} is because of random matching.

2.2 Terms of Trade

In the proceeding subsections I derive the policy functions of the agents of the model, i.e., the optimal asset holdings, their corresponding transaction costs, and the trading mechanism choices. I find that, in equilibrium, customers sort across mechanisms depending on their liquidity needs.

2.2.1 Optimal Asset Holdings and Transaction Costs

Once a customer contacts a dealer and chooses a trading mechanism, optimal asset holdings and transaction costs are set as the outcome of a Nash bargaining problem, where the dealer's bargaining powers is $\eta \in [0, 1]$ [7]. When trading on principal, execution is immediate, and so the trade surplus of the customer equals the utility gains of re-balancing positions minus the total price paid for it. On the dealer's side, her trade surplus equals the transaction cost charged minus the cost of performing principal trades. Hence, the Nash product writes

$$\{a_i^P(a), \phi_i^P(a)\} = \underset{(a', \phi')}{\arg\max} \Big\{V_i(a') - V_i(a) - p(a' - a) - \phi'\Big\}^{1-\eta} \Big\{\phi' - \theta p|a' - a|\Big\}^{\eta}.$$

[6] In terms of modeling choice, this reduced form formulation allows a link to be drawn between the demand for immediacy and dealers' inventory costs without dealing with inventories as an additional state variable. See Cohen, Kargar, Lester, and Weill (2022) for a search model with explicit inventory in OTC markets.

[7] Duffie, Gârleanu, and Pedersen (2007) model explicitly a bargaining game where agents make alternate offers. They show that the Nash bargaining powers equal the probabilities of making an offer in such a game.

The solution for optimal principal terms of trade is

$$\phi_i^P(a) = \eta\big[V_i(a_i^P(a) - V_i(a) - p(a_i^P(a) - a)\big] + (1-\eta)\big[\theta p|a_i^P(a) - a|\big], \tag{3}$$

$$a_i^P(a) = \arg\max_{a'} \quad V_i(a') - V_i(a) - p(a'-a) - \theta p|a'-a|. \tag{4}$$

The presence of inventory costs has two important consequences for principal trades. Firstly, conditional on the trade direction, inventory costs are translated into an increase (decrease) in the effective price customers pay when buying (obtain when selling). Thus, the problem becomes linear in the volume traded, and consequently, customers choose their optimal holdings independently of their current positions. Secondly, some customers might optimally not trade at all. In contrast with LR09 and the bulk of theoretical models that account for principal and agency trades, the policy function in the model allows for a no-trade region, explained by the existence of immediacy costs[8]. Whenever the gain in lifetime utility minus the inter-dealer price paid for such trade does not outweigh the immediacy costs, it is better not to trade on a principal basis. Furthermore, if keeping the current position is preferred over engaging in an agency trade, the optimal policy is not to trade at all.

These two consequences can be easily seen by optimizing equation (4) conditional on the trade direction a principal trader would pursue. Particularly, current asset holdings can be partitioned into three subsets, which I denote by $\Gamma_i \in \{Buy_i, Sell_i, NoT_i\}$:

$$\Gamma_i = \begin{cases} Buy_i: & a \,|\, [V_i(a') - a'p] - [V_i(a) - ap] > \theta p(a'-a) \quad \text{for some } a' \in (a, \infty), \\ Sell_i: & a \,|\, [V_i(a') - a'p] - [V_i(a) - ap] > \theta p(a-a') \quad \text{for some } a' \in [0, a), \\ NoT_i: & a \,|\, [V_i(a') - a'p] - [V_i(a) - ap] \leq \theta p|a'-a| \quad \forall a' \neq a. \end{cases}$$

Within each subset, optimal asset holdings can be easily characterized[9]:

$$a_i^P(a) = \begin{cases} a_i^{P,b} = \arg\max_{a'}\{V_i(a') - p(1+\theta)a'\} & \text{if } a \in Buy_i, \\ a_i^{P,s} = \arg\max_{a'}\{V_i(a') - p(1-\theta)a'\} & \text{if } a \in Sell_i, \\ a & \text{if } a \in NoT_i, \end{cases}$$

In turn, agency trades imply an expected execution delay, during which the customer might suffer a preference shock. Hence, a specific timing assumption regarding when optimal holdings and transaction

[8] Given that most of the databases are based on transaction data, the empirical evidence related to no trades is hard to find. Hendershott, Li, Livdan, and Schürhoff (2020) provide evidence of no trading in the CLO market. The authors compute a no-trading rate that goes from 7% to 30%, decreasing in the seniority tranche of the security. The CLO market features, in which trading is done through auctions and where sellers choose when to contact dealers, prevent us from reading these numbers through the lens of the present model.

[9] If the value function is increasing and strictly concave in asset holdings, these subsets are convex and the maximizers are unique. I check numerically both the convexity of the sets as well as the uniqueness of the maximizers and they both hold robustly.

costs are set is needed. In this regard, it is assumed that transaction costs are arranged when customers and dealers meet, and that optimal holdings are decided at execution. The implications of this assumption are twofold. Firstly, the model allows for order cancellation, a common practice when trading securities (Foucault, Pagano, and Röell, 2013). Secondly, agency transaction costs are set based on the expected gains from trade of customers who may suffer preference shocks while waiting[10].

A customer's expected agency trade surplus is composed by two terms. The first component is her expected utility derived from holding her current position while waiting for execution. The second component is her expected future gains from re-balancing her position. On the dealers' side, their trade surplus is just the discounted transaction cost collected. Terms of trade when agency is chosen are set according to

$$\begin{aligned}
&\{\{a_i^A\}_{i=1}^I, \phi_{i(t)}^A(a)\} \\
&= \underset{\{a_i''\}_{i=1}^I, \phi''}{\arg\max} \Big\{ \mathbb{E}_{i(t)} \Big[\int_t^{T_\beta} e^{-r(s-t)} u_{i(s)}(a) ds + e^{-r(T_\beta - t)} \big[V_{i(T_\beta)}(a''_{i(T_\beta)}) - p(a''_{i(T_\beta)} - a) - \phi'' \big] \Big] \\
&\qquad\qquad - V_{i(t)}(a) \Big\}^{1-\eta} \Big\{ \mathbb{E}_t \big[e^{-r(T_\beta - t)} \phi'' \big] \Big\}^{\eta}.
\end{aligned}$$

The optimal terms in the agency trade are

$$\begin{aligned}
\mathbb{E}_t[e^{-r(T_\beta - t)}] \phi_{i(t)}^A(a) = \eta \Big\{ \mathbb{E}_{i(t)} \Big[\int_t^{T_\beta} & e^{-r(s-t)} u_{i(s)}(a) ds \\
& + e^{-r[T_\beta - t]} \big[V_{i(T_\beta)}(a^A_{i(T_\beta)}) - p(a^A_{i(T_\beta)} - a) \big] \Big] - V_{i(t)}(a) \Big\}, \qquad (5)
\end{aligned}$$

$$a_i^A = \underset{a''}{\arg\max} \quad \{V_i(a'') - p a''\}. \qquad (6)$$

With these results at hand, I manipulate the Bellman equation (1) to reach a simpler and more intuitive representation. First, I plug in the bargaining outcomes and note that the problem is equivalent to the one faced by a customer with maximum bargaining power but smaller contact rate $\kappa = \alpha(1-\eta)$. I refer to κ as the bargaining-adjusted contact rate. Second, I use analytical expressions for all the expectations

[10] An alternative modeling choice is to assume that customers and dealers commit upon contact to trade a certain optimal volume at execution. In this case, an amplification of the effect presented in LR09 would be observed, where optimal asset holdings would be partially chosen according to the type at the moment of trading and partially according to their expected flow of types. If customers opt for agency trading, they choose their positions taking into account that they might change their preferences both before and after the execution of the trade, so the expected flow of types weight will be larger. This assumption not only is at odds with order cancellation in practice but also implies a modeling disadvantage. In particular, it requires tracking the committed trade amount within the "waiting for execution" state, adding another state variable to an already large state-space. Another alternative is to assume that the optimal volume traded and transaction costs are decided at execution. In that case, the utility that the agent loses from not having an optimal position during the waiting time would be a sunk cost and it would not be considered in the bargaining process or in the consequent terms of trade.

related to the shocks of the model.[11]

$$V_i(a) = \bar{U}_i^{\kappa}(a) + \hat{\kappa}\big[[1-\hat{\delta}^{\kappa}]\max\left\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a|\,,\, \bar{U}_i^{\beta}(a) + \hat{\beta}[\bar{V}_i^A - p(\bar{a}_i^A - a)]\right\} + \hat{\delta}^{\kappa}\sum_j \pi_j \max\left\{V_j(a_j^P) - p(a_j^P - a) - \theta p|a_j^P - a|\,,\, \bar{U}_j^{\beta}(a) + \hat{\beta}[\bar{V}_j^A - p(\bar{a}_j^A - a)]\right\}\big], \quad (7)$$

where

$$\bar{U}_i^{\nu}(a) = \Big[[1-\delta^{\nu}]u_i(a) + \delta^{\nu}\sum_j \pi_j u_j(a)\Big]\frac{1}{r+\nu}$$

$$\bar{V}_i^A = [1-\delta^{\beta}]V_i(a_i^A) + \delta^{\beta}\sum_j \pi_j V_j(a_j^A) \quad , \quad \bar{a}_i^A = [1-\delta^{\beta}]a_i^A + \delta^{\beta}\sum_j \pi_j a_j^A$$

$$\hat{\kappa} = \frac{\kappa}{r+\kappa} \quad , \quad \hat{\beta} = \frac{\beta}{r+\beta} \quad , \quad \hat{\delta^{\nu}} = \frac{\delta}{r+\delta+\nu} \quad , \quad \nu = \{\kappa, \beta\}.$$

The first term of equation (7), $\bar{U}_i^{\kappa}(a)$, is the expected utility of holding assets a until the next (bargaining-adjusted) contact with a dealer. While waiting for this contact, a customer might change her preferences, and so this term is a convex combination of the utility under the current and the future expected type. Hence, when the customer contacts a dealer she might be in two different situations: she might have avoided the preference shock or she might have received it. The corresponding probabilities of these scenarios are $(1-\hat{\delta}^{\kappa})$ and $\hat{\delta}^{\kappa}$, respectively.

If customers choose to trade on principal, the execution is immediate. The premium paid for such immediacy is expressed in a higher effective price for buyers, $p(1+\theta)$, and a lower effective price for sellers, $p(1-\theta)$. Conversely, if an agency trade is chosen, customers need to wait for execution. This waiting stage is reflected in $\bar{U}_i^{\beta}(a)$, the utility that a customer with current preference i holding asset a expects to derive until executing her agency trade. At the moment of execution, her preference may have changed, and so her expected value function, $\bar{V}_i^A$, is a convex combination across the preference space.

Equation (7) highlights the two differences between trading mechanisms. The first one is the expected execution delay that agency trading implies. The second one is the less favorable trading terms that customers face under principal trading, given the partial translation of dealers' inventory costs. These two differences define the trade-off that customers will have to solve.

2.2.2 Trading Mechanism Choice

When customers contact dealers, they must choose between an immediate principal or a delayed agency trade. I start by looking for the preference-specific current asset holding thresholds that make each customer indifferent among trading mechanisms. The indifference condition for a type i customer is given by:

[11] See the Appendix A.2 and A.3 for a step-by-step computation.

$$\left[V_i(a_i^P) - V_i(a)\right] - p(a_i^P - a) - \theta p|a_i^P - a| = \left[\bar{U}_i^{\beta}(a) + \hat{\beta}\bar{V}_i^A - V_i(a)\right] - \hat{\beta}p(\bar{a}_i^A - a), \tag{8}$$

This equation compares the trade surplus in each mechanism, which are functions of customers' difference between their current and their optimal asset holdings. To gain intuition, Figure 2 graphs, for a mid-preference customer, these trade surpluses.

Figure 2 presents two salient features. First, as current and optimal asset holdings get closer, the principal surplus goes to zero but the agency surplus remains at a positive level. This is explained because principal trading is immediate, whereas agency trading is delayed. When a customer holds the optimal principal position given her current preference, a_i^P, trading on principal would represent no surplus: the optimal position is already achieved. However, when a customer holds the optimal agency position according to her current preferences, a_i^A, trading on agency might still represent a positive expected surplus. This is because, while customers wait for execution, her preferences might change making her current position no longer optimal.

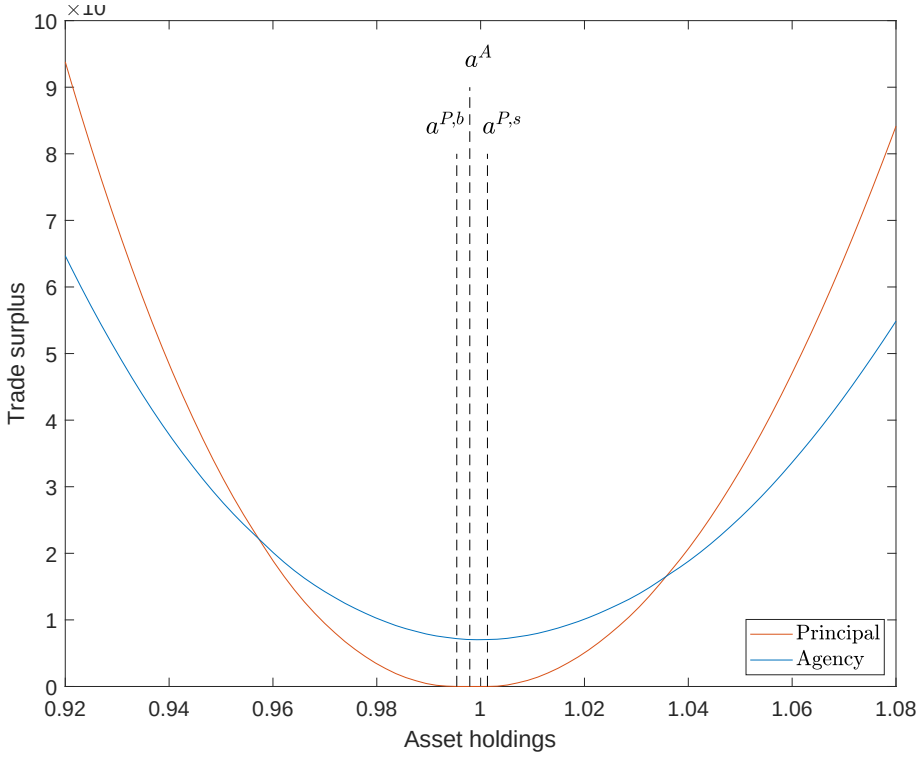
Second, customers with a larger distance between current and optimal asset holdings trade on principal. To analyze this pattern, let me consider a customer who compares whether to buy on principal or to engage in the agency trade. To further simplify the exposition, consider the limiting case where preference shocks arrive with a Poisson intensity close to zero, thus $\bar{U}_i^{\beta}(a) + \hat{\beta}\bar{V}_i^A = \frac{u_i(a)+\beta V_i(a_i^A)}{r+\beta}$ and $\bar{a}_i^A = a_i^A$. Equation 8 can be written:

$$\underbrace{\left[\frac{rV_i(a_i^A) - u_i(a)}{r+\beta}\right]}_{\text{cost of delay}} = \underbrace{p(1+\theta-\hat{\beta})(a_i^A - a)}_{\text{effective price diff}} + \underbrace{[V_i(a_i^A) - pa_i^A] - [V_i(a_i^P) - pa_i^P]}_{\text{gains from trade diff}} - \underbrace{p\theta(a_i^A - a_i^P)}_{\text{adjustment}}$$

The LHS expresses the cost of performing agency trades: while waiting for a suitable counterparty the customer will hold an unwanted position. The RHS expresses the benefits of performing agency trades. It is composed of three terms. First, agency trading allows avoiding inventory costs, and so the effective price paid is lower. Second, given that the effective price of trading on agency is more convenient than that of principal trading, a customer would trade a larger quantity in the former mechanism than in the latter. Finally, the transaction cost difference needs to be adjusted for the fact that, if the customer had traded on principal, she would have bought a smaller quantity, hence the total transaction cost difference paid to dealers would have been smaller.

The comparison between the costs and benefits of trading on agency tells us why customers with larger trading needs choose principal trades. Given a customer's preference type, only the first terms of both sides of the equation are affected by her current asset holdings. As the distance between current and optimal

Figure 2: Trading mechanism choice.



Note: This figure depicts the trade surplus under the two trading mechanisms, for a customer with preference type at the center of the distribution. The optimal asset holdings under the principal trade, for buyers and sellers, are graphed in dashed lines. The values correspond to the baseline calibration presented in section 5.3

asset holdings increases, the cost of delaying the execution increases at a faster rate than the savings given by the effective price difference. This is because the cost of each extra unit away from the optimal position is marginally increasing (utility is strictly concave), whereas the effective price difference is constant.[12]

I summarize the optimal trading mechanism rule for a customer with preference i and asset holdings a using the asset holding subset $\{\Gamma_i^P, \Gamma_i^A\}_{i=1}^I$. These are partitions of the subsets $\Gamma_i = \{Buy_i, Sell_i, NoT_i\}$, which in turn defined what the optimal trading direction was for a customer trading on principal. This decision follows from the fact that the indifference equation (8) considers the optimal asset position in each mechanism and that the principal optimal position changes with the trade direction, as it was explained in subsection 2.2. In Appendix A.4 I provide a discussion of how these sets are built.

[12]Note that, if preference shocks arrive at a positive rate, the logic follows: customers compare the costs of a delayed execution and the accumulated savings from the difference in effective prices, both terms only being affected by her current asset holdings.

2.3 Steady-state Distribution and Market Clearing

In this subsection I derive the general equilibrium steady-state equations of the model. As previously stated, a customer can be fully characterized by the triplet $\{a, i, \omega\}$. Thus, I first develop the equations needed to compute the steady-state distribution $H(a, i, \omega)$ over such individual states. Second, I state the market clearing condition to solve for the steady-state equilibrium price p.

Given that the model allows for the possibility of optimally not trading, potentially any initial asset holding $a \in R_+$ might be included in the ergodic set. In such a case, the steady-state equilibrium will be conditioned by the initial holdings of assets across customers. In order to prevent such a pathological case, I focus on calibrations where $\cap_{i=1}^{I} NoT_i^P = \emptyset$. In other words, I focus on equilibria where there is no asset position such that every type decides not to trade when holding it[13]. Under this restriction, given that $\pi_i > 0 \, \forall i$, every customer with any asset holdings will eventually trade. Hence, in the steady state, a customer will hold assets $a \in \mathcal{A}^*$, where $\mathcal{A}^* = \cup_{i=1}^{I} \{a_i^{P,b}, a_i^{P,s}, a_i^A\}$, and the steady-state distribution is characterized by the vector $n_{[a,i,\omega]}$. Equations (4) and (6) provide the optimal asset position in each kind of trade, and subsets $\{\Gamma_i^P, \Gamma_i^A\}_{i=1}^{I}$, with $\Gamma = \{Buy, Sell, NoT\}$, indicate which kind of trade customers wish to perform. These policy functions and the three shocks present in the model indicate how to track customers across the discrete state space. Since, in the steady state, the flow of customers entering and exiting each individual state should be equal, the following set of inflow-outflow equations computes the stationary distribution of the model.

$$n_{[a_i^{P,b},i,\omega_1]}: \quad \delta\pi_i \sum_{j\neq i} n_{[a_i^{P,b},j,\omega_1]} + \alpha \sum_{a\in Buy_i^P} n_{[a,i,\omega_1]} = n_{[a_i^{P,b},i,\omega_1]}\big(\delta(1-\pi_i) + \alpha \mathbf{1}_{[a_i^{P,b}\notin NoT_i^P]}\big) \tag{9}$$

$$n_{[a_i^{P,s},i,\omega_1]}: \quad \delta\pi_i \sum_{j\neq i} n_{[a_i^{P,s},j,\omega_1]} + \alpha \sum_{a\in Sell_i^P} n_{[a,i,\omega_1]} = n_{[a_i^{P,s},i,\omega_1]}\big(\delta(1-\pi_i) + \alpha \mathbf{1}_{[a_i^{P,s}\notin NoT_i^P]}\big) \tag{10}$$

$$n_{[a_i^{A},i,\omega_1]}: \quad \delta\pi_i \sum_{j\neq i} n_{[a_i^{A},j,\omega_1]} + \beta \sum_{a\in \mathcal{A}^*} n_{[a,i,\omega_2]} = n_{[a_i^{A},i,\omega_1]}\big(\delta(1-\pi_i) + \alpha \mathbf{1}_{[a_i^{A}\notin NoT_i^P]}\big) \tag{11}$$

$$n_{[a,i,\omega_1]}: \quad \delta\pi_i \sum_{j\neq i} n_{[a,j,\omega_1]} = n_{[a,i,\omega_1]}\big(\delta(1-\pi_i) + \alpha \mathbf{1}_{[a\notin NoT_i^P]}\big), \quad a \in \cup_{j\neq i}\{a_j^{P,b}, a_j^{P,s}, a_j^A\} \tag{12}$$

$$n_{[a,i,\omega_2]}: \quad \delta\pi_i \sum_{j\neq i} n_{[a,j,\omega_2]} + \alpha n_{[a,i,\omega_1]} \mathbf{1}_{[a\in\Gamma_i^A]} = n_{[a,i,\omega_2]}\big(\delta(1-\pi_i) + \beta\big), \quad a \in \mathcal{A}^* \tag{13}$$

The left-hand side of these equations represents the inflow in a specific individual state, and the right-hand side represents the outflow. As Figure 1 shows, in any time interval, three kinds of forces might move customers across states. Let us first consider the preference shock. The mass of customers of an individual state with preference i increases whenever customers from other states, with the same asset holdings and in the same waiting stage, receive the preference shock i. This happens with Poisson intensity $\delta\pi_i$. Similarly,

[13] As will be explained in section 5, the GMM procedure used to estimate the model searches through the parametric space in an unrestricted manner, yielding a calibration where the restriction here imposed is not binding

that mass of customers decreases whenever customers therein are hit by preference shocks other than i. This happens with intensity $\delta(1-\pi_i)$. Second, let us consider the contact with dealer shock. This shock is received only by people waiting for a dealer, i.e., by customers within states where $\omega = \omega_1$, and happens with intensity α. Customers with current asset holdings that make them want to buy (sell) on principal will flow towards the state in which optimal asset holdings for principal buyers (sellers) correspond with their preference type. On the contrary, a customer with current asset holdings such that she opts for an agency trade will flow towards the waiting-for-execution stage, i.e., $\omega = \omega_2$, keeping both her holdings and preference type. It is worth noting that not all customers hit by this shock would travel across the state space. If a customer chooses not to trade, then she will remain in her current state until a preference shock eventually hits her. Finally, the execution shock, which happens with intensity β, moves customers across waiting stages. Obviously, such shock is received only by customers waiting for the execution of their trades, i.e., in states where $\omega = \omega_2$. Once a customer gets her agency trade executed, she goes back to the "waiting for dealers" stage. Since customers decide on optimal holdings at the moment of execution, this shock will move customers toward the state in which optimal agency asset holdings correspond with their preference type.

The set of equations (9)-(13) can be represented by a transition matrix $T_{[3I\times I\times 2]}$, with attached transition probabilities $\pi^T_{n,n'}$, which denote the probability of moving from a state n towards a state n' in a given time length. Such a transition matrix can be used to update the vector of individual states masses until reaching the unique limit invariant distribution $n = \lim_{k\to\infty} n_0 T^k$, where n_0 is any initial distribution. Th.11.4 in Stokey, Lucas, and Prescott (1989) provides the conditions for this convergence result [14]. Once solved for the stationary distribution, the market clearing equation can be computed, and thus the steady-state equilibrium price p can be found. Aggregate gross demand in this secondary market is given by the weighted sum of individual states demands. Aggregate gross supply, in turn, is fixed by A. Therefore, the equilibrium price is the one at which the following market clearing equation holds:

$$\sum_{h=1}^{2}\sum_{i=1}^{I}\sum_{a\in\mathcal{A}^*} a n_{[a,i,\omega_h]} = A. \tag{14}$$

Note that, in the steady state, trading occurs constantly but the aggregate asset position is held constant. Given that all trades are cleared in the inter-dealer market, the market clearing condition (14) implies that the inter-dealer market is at equilibrium at all times. Of course, our steady state allows for a

[14] Basically, there should exist at least one state that receives inflows from all states with strictly positive probability. A sufficient condition for this to happen is that there exists a type i and a type j such that $\mathcal{A}^*_i \in Buy^P_j$, $\mathcal{A}^*_i \in Sell^P_j$ or $\mathcal{A}^*_i \in Buy^A_j \cap Sell^A_j$, where $\mathcal{A}^*_i = [a_i^{P,b}, a_i^{P,s}, a_i^A]$. Firstly, $\pi_i > 0\,\forall i$ and $\delta \in (0,1)$; therefore all types can turn into type i. Secondly, after customers of type i execute their trades, they go back to the waiting for a dealer stage. Finally, the condition described guarantees that, when those customers contact a dealer with their preferences i intact, they choose the same trading mechanism and eventually obtain the same optimal asset position. Thus such latter individual state would receive inflows directly or indirectly from all individual states. I check numerically and this condition robustly holds.

situation where the excess of demand in one mechanism is compensated by an excess of supply in the other.

2.4 Equilibrium

An equilibrium for this model is defined as a list of optimal asset holdings $\{a_i^P(a), a_i^A\}_{i=1}^I$, transaction costs $\{\phi_i^P(a), \phi_i^A(a)\}_{i=1}^I$, trading mechanism sets $\{\Gamma_i^P, \Gamma_i^A\}_{i=1}^I$ where $\Gamma = \{Buy, Sell, NoT\}$, stationary distribution $n_{[a,i,\omega]}$ and price p such that $\{a_i^P(a), a_i^A\}_{i=1}^I$ satisfies (4) and (6), $\{\phi_i^P(a), \phi_i^A(a)\}_{i=1}^I$ satisfies (3) and (5), $\{\Gamma_i^P, \Gamma_i^A\}_{i=1}^I$ are defined using thresholds satisfying (8), $n_{[a,i,\omega]}$ satisfies (9)-(13), and p satisfies (14).

In contrast with LR09, where the equilibrium can be found analytically, the model here presented needs to be solved numerically. The main difference with respect to LR09 in this regard is that current asset holdings affect not just the optimal portfolio, but also the trading mechanism chosen. To solve for the steady state of the model for any given inter-dealer price, p, I rely on the value function iteration method, enhanced with Howard's improvement step [15]. This procedure returns the policy and value functions conditional on p. In turn, these functions are nested within the computation of equation 14, which solves the inter-dealer price that clears the market in the steady state. The algorithm is described in detail in Appendix A.6.

3 Equilibrium Allocations

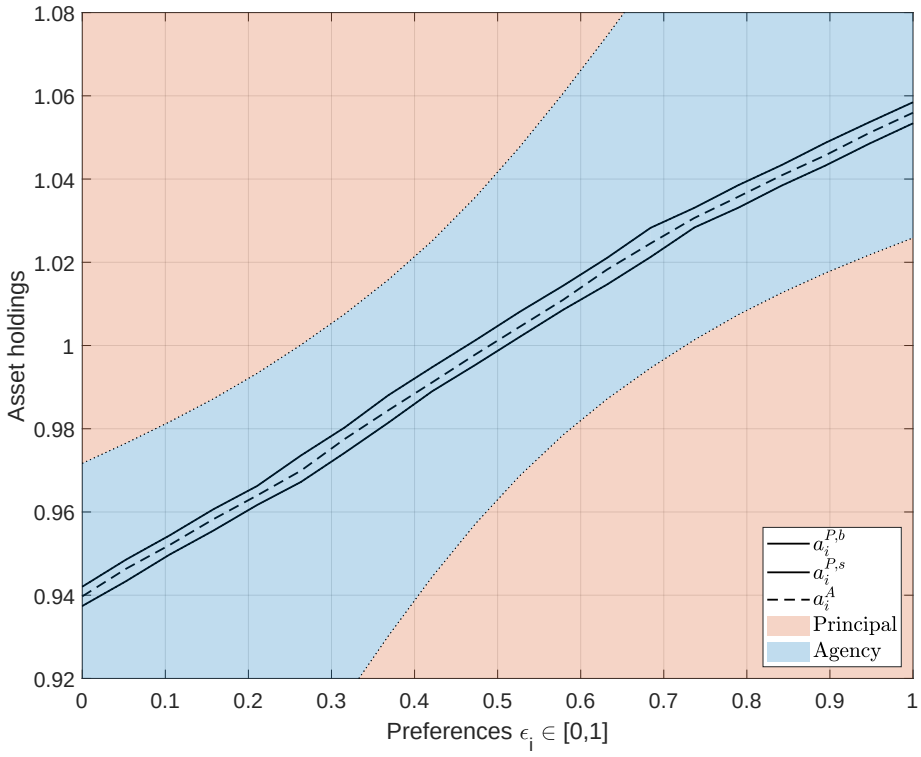
In this section I study numerically the policy functions of the model. I use the parameter values that will be estimated in section 5. I initially map customers' preferences and current asset holdings with their optimal asset holdings and mechanism choices. I show that customers sort themselves across trading mechanisms according to their trading needs. After characterizing the pool of trades in each mechanism, I describe how such characteristics are translated into the transaction costs customers pay.

3.1 Equilibrium Asset Holdings and Trading Mechanism

The policy functions are presented in Figure 3. For each asset holding and preference type pair, $\{a, i\}$, I compute both the optimal asset holdings conditional on the trading mechanism and the trading mechanism choice. Regarding the optimal asset holdings, the lower and upper solid lines represent the buyer's and seller's optimal holdings under the principal trade, $a^{P,b}$ and $a^{P,s}$, respectively. Conditional on trading on a principal basis, these two lines define three regions: a customer with assets $a < a^{P,b}$ would be a buyer, with holdings $a > a^{P,s}$ would be a seller, and with current assets $a \in [a^{P,b}, a^{P,s}]$ would not trade on principal. These three regions are a direct consequence of the inclusion of inventory costs. On the one hand, in the principal mechanism, buyers trade at an effective price higher than the one received by sellers. Hence, conditional on preference type, buyers' optimal quantity is smaller than that of sellers. On the other hand, the principal

[15] See Appendix A.5 for the necessary and sufficient conditions to use value function iteration as the solution method.

Figure 3: Optimal asset holdings and trading mechanism choice.



Note: This figure depicts the estimated model policy functions of each customer, conditional on her preference type and current holdings. The lower and upper solid lines represent the buyer's and seller's optimal asset holdings under the principal trade, $a^{P,b}$ and $a^{P,s}$, respectively. The dashed line represents the optimal asset holdings under the agency trade, a^A. Regarding the mechanism choice, the principal and agency regions are shaded in orange and blue, respectively.

trade surplus of those customers with current holdings between the buyer's and seller's optimal holdings is smaller than the principal costs faced by the dealers. Hence, there are no gains from trade and those customers decide not to trade on a principal basis. The agency optimal holdings, in turn, are represented by the dashed black line a^A. These positions are between those of the principal buyers and the principal sellers. Recall that agency trading does not imply any cost for dealers. Since dealers face no costs, the transaction cost charged to customers, conditional on trading volume, is smaller. The direct consequence is that the effective agency price is between the effective principal buy and sell prices, and thus agency optimal holdings are between those of the principal traders.

Figure 3 also presents the trading mechanism each customer chooses. The blue shaded area represents the agency region: customers who decided to wait for execution instead of paying the cost for immediacy or waiting to contact another dealer. As can be seen, in the estimated model (see section 5), every potential principal non-trader, i.e., customers with holdings $a \in [a^{P,b}, a^{P,s}]$, finds that engaging in an agency trade is better than not trading at all and waiting for a new contact with a dealer. Finally, the orange shaded area

stands for customers that trade on principal.

To better understand these policy functions, consider for example customers with preferences $\epsilon_i = 0.4$. When contacted by a dealer, these customers compute their optimal asset position as principal traders, $a_i^{P,b}$ or $a_i^{P,s}$, and their expected optimal position after the waiting period of the agency trade, $\bar{a}_i^A$. Given these optimal asset positions, they evaluate, using eq (8), which trade to perform. As Figure 3 shows, customers owning roughly less than 0.94 units of the asset perform a principal buy. Customers holding between 0.94 and 1.02 units perform an agency trade. Finally, customers holding assets above 1.02 choose to sell on a principal basis.

Figure 3 confirms an earlier observation: principal traders are concentrated in the extremes of the preference-assets state space. Firstly, conditional on preference types, principal trading is mostly performed by customers with current asset holdings far away from their optimal ones. As it was discussed in subsection 2.2.2, this is because the utility loss of each extra unit away from the optimal position is marginally increasing, whereas the principal premium that needs to be paid to avoid such costs is constant. Secondly, conditional on current asset holdings, agency trading is mostly performed by customers with preferences close to the mean. This is because optimal asset positions are increasing in preference types: customers with extreme preferences will find themselves more often far away from their optimal position than customers with moderate preferences. Given the relation between trading mechanism choice and the distance between the current and optimal position, the model tells us that customers with moderate preferences are more likely to perform agency trades, while customers with extreme preferences are more likely to trade on principal.

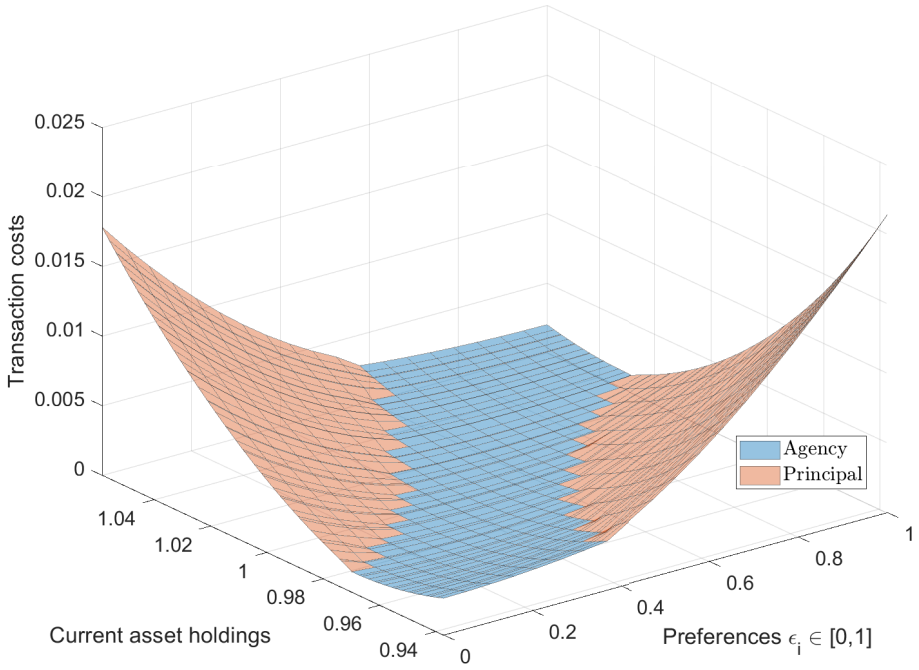
3.2 Equilibrium Transaction Costs

I next present the distribution of transaction costs paid by customers. As equations (3) and (5) show, these costs are solved through Nash bargaining; therefore, they incorporate the specific characteristics of the trade. Particularly, transaction costs are convex combinations between customers' expected trading surplus and dealers' inventory cost. In turn, these objects are functions of the asset holdings and preference held by the customer when she contacts the dealer, and of the resulting trading mechanism chosen. Figure 4, which maps transaction costs with the asset-preference state-space, depicts such heterogeneity.

Overall, the broad features of transaction costs in LR09 still hold. For example, marginal transaction costs are increasing in the traded volume. A marginally decreasing utility implies that, given a certain optimal position the marginal trading surplus is increasing in the volume traded. The bargaining protocol used implies that transaction costs are linear functions of such surpluses; thus, they inherit the property [16]. On top of this, two interesting properties regarding the trading mechanism distinction are observed.

[16] Pinter et al. (2022) study the relation between trading costs and trading size in the UK government and corporate bond markets. In contrast with other empirical papers on the topic, their database has both customers' and dealers' identities. This feature allows them to control for customer cross-section variation when computing the trade size effect. In line with the model here developed, they show that, conditional on the customer's identity, trading costs are increasing in trade size.

Figure 4: Transaction costs under each trading mechanism.



Note: This figure depicts the estimated model transaction cost paid by each customer, conditional on her preference type and current asset holdings. The orange-shaded area refers to principal costs. The blue shaded area refers to (present valued) agency costs.

Firstly, principal transaction costs are on average larger than those of agency trades. On the one hand, principal traders exchange larger quantities and thus obtain larger trade surpluses. On the other hand, even conditioning on the customer's trading surplus, principal transaction costs are still larger than agency, given the inclusion of the translated inventory costs. This later feature is evident from the presence of jumps at the thresholds [17]. Secondly, principal transaction costs increase at a higher rate when moving both towards extreme preferences and towards larger trading quantities. When customers trade on agency, they are subject to preference shocks. This implies that agency customers anticipate that both the utility they get from current holdings and the optimal trading volume may change while waiting for execution. Hence,

[17] If current asset holdings equal asset thresholds, the indifference condition (8) indicates that the net trade surplus for any preference type under both mechanisms is the same. At such current asset holdings, from the definition of inventory costs and as long as asset holding thresholds and principal optimal holdings are different, inventory costs will be positive. Given that transaction costs are convex combinations of customers' trade surpluses and dealer costs, at the thresholds principal costs exceed (present valued) agency costs exactly by the inventory costs amount.

$$\phi_i^P(\hat{a}_i) - \theta p|a_i^P - \hat{a}_i| = \hat{\beta}\phi_i^A(\hat{a}_i)$$

This result can be easily obtained combining equations (3), (5), and (8).

instead of the certain immediate trade surplus given by principal trades, agency customers need to consider an average surplus based on expected preference shocks. Therefore, across the agency region expected trade surpluses, and consequently transaction costs, are relatively flatter [18].

As can be seen, the model yields a rich heterogeneity both across and within trading mechanisms. Customers with large (small) trading needs and holding relatively extreme (moderate) preference types choose principal (agency) trades. Accordingly, those customers trading on principal pay an average higher transaction costs than those trading on agency. Finally, given the possibility of changing preferences while waiting for execution, transaction costs are relatively flatter across the state-space within the agency region. These differences will play a key role when addressing composition effects. If the customers that migrate across trading mechanisms when market conditions change paid different costs than the non-migrating ones, then the samples over which transaction costs pre and post-change are measured will not be comparable. The next section computes average transaction costs as empirical researchers would and develops a strategy to control for such change in samples.

4 Liquidity Measures

Recent empirical literature on OTC markets argues that liquidity conditions have changed during the last decade. In particular, researchers document a shift in trading volume, from immediate principal towards delayed agency trades, accompanied by an increase in immediacy costs (Anderson and Stulz, 2017; Schultz, 2017; Bao, O'Hara, and Zhou, 2018; Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; Dick-Nielsen and Rossi, 2019; O'Hara and Zhou, 2021; Kargar, Lester, Lindsay, Liu, Weill, and Zúñiga, 2021; Choi, Huh, and Shin, 2023). In this section I compute the model's liquidity measures necessary to understand and analyze this phenomenon. Firstly, I compute the turnover rate and average transaction costs that serve as theoretical counterparts of the empirical measures. Secondly, I build counterfactual measures of transaction costs which account for composition effects. By comparing average and counterfactual measures I obtain the size and sign of the bias. These objects are used to revisit how liquidity changes when there are higher regulatory costs or when the speed of execution of agency trades increases.

4.1 Turnover and Transaction costs

To compute liquidity measures, it is useful to regroup the optimal trading mechanism sets. Define $P_i \equiv Buy_i^P \cup Sell_i^P$, $A_i \equiv Buy_i^A \cup Sell_i^A \cup NoT_i^A$, and $NT_i \equiv NoT_i^P$, as the sets under which customers of preference i trade on principal, on agency, or do not trade at contact with dealers. The turnover rate is computed as the ratio between the total dealer-customer volume traded per unit of time and the aggregate

[18] In Appendix A.7 I graph transaction costs per dollar traded. All the features previously mentioned hold if this alternative specification is considered.

asset supply. The supply of assets is fixed at A, so I only need to compute the volume. Principal trades are performed by customers who are waiting to contact a dealer and prefer immediate trades, i.e., customers in state $n_{[a,i,\omega_1]}$, where $a \in P_i$. These contacts happen at rate α, and the volume traded in each transaction is $|a_i^P(a) - a|$. In turn, agency trades are performed by customers who had already agreed to conduct such contract and therefore are waiting for its execution. These customers are found in states $n_{[a,i,\omega_2]}$, where $a \in \mathcal{A}^*$. They execute their contracts at rate β, and exchange volume according to $|a_i^A - a|$. The turnover in each mechanism, expressed in percentage points, is:

$$\mathcal{T}^P = 100 \times \frac{1}{A}\alpha \sum_{i\in\mathcal{I}} \sum_{a\in P_i} n_{[a,i,\omega_1]} |a_i^P - a|, \tag{15}$$

$$\mathcal{T}^A = 100 \times \frac{1}{A}\beta \sum_{i\in\mathcal{I}} \sum_{a\in\mathcal{A}^*} n_{[a,i,\omega_2]} |a_i^A - a|. \tag{16}$$

The aggregated turnover is just the sum of the turnovers in both mechanisms, $\mathcal{T} = \mathcal{T}^P + \mathcal{T}^A$. In a similar fashion, the volume-weighted average transaction costs for each trading mechanism can be computed. To do this, I first compute the transaction cost per (*numeraire*) dollar traded. Then these figures are averaged using the total volume share of each contract as weights. A consideration must be made regarding the computation of per-dollar costs for agency trades. In such contracts, transaction costs are arranged at contact with dealers and the optimal asset positions are chosen at execution. While waiting for execution, customers can suffer preference shocks. Hence, two customers with the same agency contract might end up trading different volumes. Hence, I compute the aggregated volume for each contract. To do so, I rely on the Law of Large Numbers and track customers across the state-space while they are waiting for execution. The weighted average transaction cost in each mechanism, expressed in basis points (bps), is:

$$\mathcal{S}^P = 10000 \times \sum_{i\in\mathcal{I}} \sum_{a\in P_i} \frac{n_{[a,i,\omega_1]}|a_i^P - a|}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i} n_{[a,i,\omega_1]}|a_i^P - a|} \frac{\phi_{a,i}^P}{|a_i^P - a|p}, \tag{17}$$

$$\mathcal{S}^A = 10000 \times \sum_{i\in\mathcal{I}} \sum_{a\in A_i} \frac{n_{[a,i,\omega_1]} rav_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i} n_{[a,i,\omega_1]} rav_{a,i}} \frac{\phi_{a,i}^A}{rav_{a,i}p}. \tag{18}$$

where $rav_{a,i}$ stands for the realized agency volume for contracts signed by customers holding i preference and a assets at the moment of contact with dealers[19]:

$$rav_{a,i} = (1-\hat{\delta})|a_i^A - a| + \hat{\delta}\sum_{j\in\mathcal{I}} \pi_j |a_j^A - a|.$$

The average transaction cost unconditional on trading mechanism is just the weighted average of the

[19]Note that $rav_{a,i}$ takes into account the possibility of contracting an agency trade but ending up not trading. This happens whenever the current and optimal asset holdings are equal at execution. An alternative computation tracking agency customer until execution yields the same result.

previous figures: $\mathcal{S} = [\mathcal{T}^P \mathcal{S}^{\mathcal{P}} + \mathcal{T}^A \mathcal{S}^{\mathcal{A}}]/\mathcal{T}$. As can be seen, average transaction costs are functions of both the costs associated with each transaction and the steady-state mass of customers who endogenously trade in each mechanism. When the economy changes, these two vectors are affected. Thus, the model is able to capture not only the change in transaction cost per trade, but also the sample composition effects.

4.2 Transaction Costs Decomposition

To account for composition effects, I build counterfactual measures of average transaction costs fixing the samples over which they are measured. In order to do so, I decompose the steady-state distribution into those customers that, under alternative parametrizations, would migrate across mechanisms and those that would not. Counterfactual transaction cost measures are computed using only the subsamples of non-migrating customers.

Recall that, when customers contact dealers, they choose their optimal trading mechanism according to thresholds that satisfy the indifference condition (8). These thresholds define trading mechanism sets, i.e., preference-specific asset holding sets under which customers choose to trade on principal, on agency, or not to trade at all, P_i, A_i and NT_i, respectively. Consider firstly alternative parametrizations, denoted by q, and compute their steady-state trading mechanism sets. Secondly, for each preference type, compute the intersections across parametrizations between these trading mechanism sets. To ease the exposition, I only consider two parametrizations, $q \in \{0, 1\}$, but the method can be easily extended to account for any number of parametrizations. Table 1 presents the resulting subsets. Diagonal cells include customers that choose the same trading mechanism under the two scenarios. I call these customers non-migrants. Conversely, non-diagonal cells include customers who change their optimal mechanism when facing different scenarios. I call these customers migrants. For example, the population of customers with preference i holding assets $a \in [P^0, A^1]_i$ would trade on principal under $q = 0$ and would migrate towards agency under $q = 1$ [20].

Table 1: Sample decomposition

	P_i^1	A_i^1	NT_i^1
P_i^0	$P_i^0 \cap P_i^1$	$P_i^0 \cap A_i^1$	$P_i^0 \cap NT_i^1$
A_i^0	$A_i^0 \cap P_i^1$	$A_i^0 \cap A_i^1$	$A_i^0 \cap NT_i^1$
NT_i^0	$NT_i^0 \cap P_i^1$	$NT_i^0 \cap A_i^1$	$NT_i^0 \cap NT_i^1$

These subsets allow defining subsamples over which to compute transaction costs. To this end, I add new notation. Superscripts attached to cost measures indicate both the trading mechanism and the

[20] If $Q > 2$ number of parametrizations are considered, 3^Q number of subsets within a Q-dimension matrix are obtained. The diagonal of such higher-order matrix defines customers that choose the same trading mechanism under all the alternative parametrizations. For example, customers with preference i that remain trading on principal regardless of the parametrization used are those with assets $a \in \cap_{q=1}^{Q} P_i^q$.

parameters used. In turn, subscripts, whenever present, denote which trading subsets were used to define the subsample. For example, $\mathcal{S}^{P,0}_{P^0,P^1}$ refers to principal transaction costs paid under scenario $q = 0$ by customers who trade on principal both under $q = 0$ and $q = 1$. In turn, $w^{P,0}_{P^0,P^1}$ refers to the volume share accounted for such transactions under scenario $q = 0$. Finally, I can decompose the change in transaction costs for each mechanism due to a parametric change. Consider $q = 0$ as the initial scenario, and $q = 1$ as the new one.[21]

$$\begin{aligned}\Delta\mathcal{S}^P = \mathcal{S}^{P,1} - \mathcal{S}^{P,0} = &\underbrace{\mathcal{S}^{P,1}_{P^0,P^1} \times w^{P,1}_{P^0,P^1} - \mathcal{S}^{P,0}_{P^0,P^1} \times w^{P,0}_{P^0,P^1}}_{\text{Principal non-migrants}} \\ &+ \underbrace{\mathcal{S}^{P,1}_{A^0,P^1} \times w^{P,1}_{A^0,P^1} + \mathcal{S}^{P,1}_{NT^0,P^1} \times w^{P,1}_{NT^0,P^1}}_{\text{Inflow migration}} - \underbrace{\mathcal{S}^{P,0}_{P^0,A^1} \times w^{P,0}_{P^0,A^1} - \mathcal{S}^{P,0}_{P^0,NT^1} \times w^{P,0}_{P^0,NT^1}}_{\text{Outflow migration}},\end{aligned} \tag{19}$$

$$\begin{aligned}\Delta\mathcal{S}^A = \mathcal{S}^{A,1} - \mathcal{S}^{A,0} = &\underbrace{\mathcal{S}^{A,1}_{A^0,A^1} \times w^{A,1}_{A^0,A^1} - \mathcal{S}^{A,0}_{A^0,A^1} \times w^{A,0}_{A^0,A^1}}_{\text{Agency non-migrants}} \\ &+ \underbrace{\mathcal{S}^{A,1}_{P^0,A^1} \times w^{A,1}_{P^0,A^1} + \mathcal{S}^{A,1}_{NT^0,A^1} \times w^{A,1}_{NT^0,A^1}}_{\text{Inflow migration}} - \underbrace{\mathcal{S}^{A,0}_{A^0,P^1} \times w^{A,0}_{A^0,P^1} - \mathcal{S}^{A,0}_{A^0,NT^1} \times w^{A,0}_{A^0,NT^1}}_{\text{Outflow migration}}.\end{aligned} \tag{20}$$

The introduced decomposition highlights the interaction between the changing average costs in each subsample and the changing subsample weights. It has three components. The first term accounts for the non-migrants' effect. On the one hand, customers who keep on trading under the same mechanism may pay different costs. On the other hand, the volume share of those customers may also change. The second and third terms are related to the migrants' effect. Under a new scenario, some customers may decide to change their optimal trading strategy. Customers that represent an inflow into a given mechanism add up their costs to the overall average. Conversely, customers that imply an outflow subtract their previously paid costs from that average.

Equations (19) and (20) provide a natural way of defining counterfactual measures of transaction costs free of composition effects. If the samples within the trading mechanism were held constant, non-migrant customers would have full weight in all scenarios. Therefore, I define the composition-free measures of transaction cost under parametrization q, $\tilde{\mathcal{S}}^P(q)$ and $\tilde{\mathcal{S}}^A(q)$, as the costs measured within the non-migrant samples. In turn, the composition-free measures of transaction cost change, $\Delta\tilde{\mathcal{S}}^P$ and $\Delta\tilde{\mathcal{S}}^A$, are set to account only for such non-migrant figures. Finally, the composition effect bias measures, CE^P and CE^A,

[21] See Appendix A.8 for details.

are defined as the fraction of the change in transaction costs due to migration.

$$\tilde{\mathcal{S}}^P(q) \equiv \mathcal{S}^{P,q}_{P^0,P^1}, \tag{21}$$

$$\tilde{\mathcal{S}}^A(q) \equiv \mathcal{S}^{A,q}_{A^0,A^1}, \tag{22}$$

$$\Delta\tilde{\mathcal{S}}^P \equiv \mathcal{S}^{P,1}_{P^0,P^1} - \mathcal{S}^{P,0}_{P^0,P^1}, \tag{23}$$

$$\Delta\tilde{\mathcal{S}}^A \equiv \mathcal{S}^{A,1}_{A^0,A^1} - \mathcal{S}^{A,0}_{A^0,A^1}, \tag{24}$$

$$CE^P \equiv 1 - \Delta\tilde{\mathcal{S}}^P/\Delta\mathcal{S}^P, \tag{25}$$

$$CE^A \equiv 1 - \Delta\tilde{\mathcal{S}}^A/\Delta\mathcal{S}^A. \tag{26}$$

The introduction of composition-free measures of transaction cost changes sheds light on the necessary conditions for the existence of composition effects mentioned in the introduction of this paper. In the first place, migrating customers are needed. Their absence would imply that the samples under the two scenarios are equal. Secondly, the costs paid by migrating and non-migrating customers should be different. Otherwise, the in-flowing and out-flowing migrants would not alter the average costs of each mechanism. Finally, as long as the difference between costs paid by migrants and non-migrants is driven by unobservable characteristics, empirical estimates would include a composition effect bias. Our model suggests that such an unobservable characteristic is the idiosyncratic trading surplus of each customer, which in turn is a function of both the distance between current and optimal positions and the idiosyncratic utility each customer derives from holding the assets.

5 Estimation

In this section I bring the model to the data. Particularly, I target key moments of the US corporate bond secondary market. I initially outline the estimation method. Later I describe how to compute the moments used in such a procedure, both theoretically and empirically. Finally, I present the estimation results and the moments' variation that allows for the identification of the parameters.

5.1 Estimation Procedure

The baseline parametrization of the model will consist of a combination of externally calibrated parameters and estimated parameters. I set the unit of time to be a month. In line with recent research on structural estimation of related search models (Coen and Coen, 2022; Pinter and Uslu, 2022), I consider a monthly discount rate of 0.5%. The support of the preferences shifters ϵ_i is normalized to $\left\{\frac{i-1}{I-1}\right\}_{i=1}^{I}$, with $I = 20$. In the model, expanding or contracting the support of ϵ_i only scales up or down the nominal variables, i.e., the inter-dealer price and the transaction costs. Given that I will focus on transaction costs per (*numeraire*)

dollar traded, normalizing such support does not affect the results. Similarly, the supply of assets A only scales up and down both nominal and real variables. Since all real variables will be expressed in terms of the total asset supply, I normalize $A = 1$. As was shown in subsection 2.2.1, the bargaining power of the dealers, η, is closely related to the arrival rate of opportunities to trade, α. In a nutshell, customers are indifferent between contacting high bargaining power dealers often and low bargaining power dealers scarcely. This precludes me from disentangling these two parameters, and therefore I opt to externally calibrate the bargaining power and to estimate the contact rate with dealers. I follow Hugonnier, Lester, and Weill (2020) and set $\eta = 0.95$. Finally, the last object externally calibrated is the probability distribution assigned to each preference type. I follow Coen and Coen (2022) and assume that such preferences are uniformly distributed, $\pi_i = 1/I \ \forall i$. In the appendix C.2 I show that the main results qualitatively hold when considering lower or higher bargaining powers or alternative preference distributions.

The remaining parameters of the model are the rates at which customers contact dealers, suffer preference shocks and execute their agency trades, α, δ and β respectively, the dealer's marginal inventory costs, θ, and the utility curvature parameter, σ. I jointly estimate these parameters using generalized method of moments (GMM). Particularly, I define the vector $\upsilon = [\alpha, \delta, \beta, \theta, \sigma]$ and estimate $\hat{\upsilon}$ as the argument that minimizes the percentage difference between the implied theoretical moments, $m(\upsilon)$, and the computed empirical moments, m_s:

$$\hat{\upsilon} = \underset{\upsilon \in \Upsilon}{\arg\min} [(m(\upsilon) - m_s) \oslash m_s]' W [(m(\upsilon) - m_s) \oslash m_s],$$

where $\oslash$ is element-wise division. Note that by using percentage deviation I ensure that the scales of the different moments do not play any role in the procedure. In line with the literature, W is set as the identity matrix, thus assigning equal weights to the different moments (Coen and Coen, 2022; Pinter and Uslu, 2022).

5.2 Moments

I choose a set of moments that covers both quantities and prices, as well as the interaction among them. I target the overall monthly turnover, $\mathcal{T}$, the volume weighted average transaction costs in each mechanism, S^P and S^A, and the slopes of the transaction costs over the trade size, for each mechanism, γ^P and γ^A. In particular, to gauge the size of composition effects, it is fundamental to target the differential transaction costs paid by migrants and non-migrants. Section 6 will show that migrants are located in the extremes of the trading size distribution, conditional on preference type. Thus matching the slope of transaction costs on trading size, γ^P and γ^A, informs about the differential transaction costs paid by migrants and non-migrants. In subsection 5.3 I discuss how the variation of these moments can identify the vector of parameters υ.

5.2.1 Theoretical Moments

For any given vector υ, I compute the theoretical moments using the steady-state equilibrium of the model. These are:

- Monthly turnover:

$$\mathcal{T} = 100 \times \frac{\alpha \sum_{i\in\mathcal{I}} \sum_{a\in P_i} n_{[a,i,\omega_1]} |a_i^P - a| + \beta \sum_{i\in\mathcal{I}} \sum_{a\in\mathcal{A}^*} n_{[a,i,\omega_2]} |a_i^A - a|}{A}, \tag{M.1}$$

- Volume weighted average transaction cost in each mechanism:

$$\mathcal{S}^P = 10000 \times \sum_{i\in\mathcal{I}} \sum_{a\in P_i} \frac{n_{[a,i,\omega_1]} |a_i^P - a|}{\sum_{i\in\mathcal{I}} \sum_{a\in P_i} n_{[a,i,\omega_1]} |a_i^P - a|} \frac{\phi_{a,i}^P}{|a_i^P - a|p}, \tag{M.2}$$

$$\mathcal{S}^A = 10000 \times \sum_{i\in\mathcal{I}} \sum_{a\in A_i} \frac{n_{[a,i,\omega_1]} rav_{a,i}}{\sum_{i\in\mathcal{I}} \sum_{a\in A_i} n_{[a,i,\omega_1]} rav_{a,i}} \frac{\phi_{a,i}^A}{rav_{a,i}p}. \tag{M.3}$$

- Transaction cost - trade size slope in each mechanism:

$$\gamma^P = 100 \times \frac{cov(\phi^P/(|a^P - a|p), |a^P - a|)}{var(|a^P - a|)}, \tag{M.4}$$

$$\gamma^A = 100 \times \frac{cov(\phi^A/(rav \times p), rav)}{var(rav)} \tag{M.5}$$

where the variance and covariance equations are described in the Appendix B.1.

5.2.2 Empirical Moments

To compute the empirical moments, I rely on transaction data of the US corporate bond secondary market, from January 2016 to December 2019. Specifically, I use the academic Trade Reporting and Compliance Engine (TRACE) database, produced by the Financial Industry Regulatory Authority (FINRA).

Given the well-known presence of reporting errors, the data is filtered following the procedure outlined in Dick-Nielsen and Poulsen (2019) [22]. I also remove the duplicated inter-dealer trades and those trades in which dealers transfer bonds to their non-FINRA affiliates for book-keeping purposes (Adrian, Boyarchenko, and Shachar, 2017) [23]. I further merge this transaction-level data with bond-level variables from the Mergent Fixed Income Securities Database (FISD). Following the empirical literature, several filters are applied (e.g., Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; Friewald and Nagler, 2019; Kargar, Lester, Lindsay, Liu, Weill, and Zúñiga, 2021). Among them, the most significant are dropping bonds that are

[22] Both the algorithm and the filter results can be downloaded from my personal website.
[23] Starting on November 2, 2015, FINRA provides explicit labels for the so-called book-keeping trades.

preferred, convertible or exchangeable, yankee bonds, bonds with a sinking fund provision, variable coupon, with time to maturity of less than a year, or issued less than two months before the transaction date[24].

Needless to say, the empirical transaction costs are partially driven by features not present in my model, e.g., default risk and asymmetric information. In that regard, to improve the likelihood of my model capturing the targeted moments I exclude from the sample those bonds that had been labeled as high yield at any point during my sample period [25].

One important feature of the academic version of TRACE is that it contains anonymous identities for each dealer. I exploit that feature to identify principal and agency trades. The idea underlying the identification is that the shorter the time it takes for a dealer to offload a position, the bigger it is the probability that those trades had been previously arranged and thus intermediated on an agency basis (Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; Kargar, Lester, Lindsay, Liu, Weill, and Zúñiga, 2021; Choi, Huh, and Shin, 2023). I classify customer-dealer trades into three categories: those that are quickly offset with other customers, those that are quickly offset with other dealers, and those that are not offset. The first and third categories are agency and principal trades, respectively. Specifically, for each customer-dealer trade, I look for all the offsetting trades of the same dealer in the same bond, within a 15-minute window. If at least 50% of its volume was offset, and the majority of such volume was offset with customers, I label it as an agency trade. If less than 50% of its volume was offset, I label it as a principal trade. The remaining transactions are disregarded.

Two subtleties about the principal/agency distinction are worth noting. First, this procedure allows for multiple matching, in the sense that a single trade can be offset by several trades of the opposite direction. Second, the algorithm may encounter competing trades. In such case I form pairs with the trades that are closer in time firstly, and closer in volume secondly [26].

Once the data has been filtered and only principal or agency customer-dealer trades are kept, I proceed to compute the empirical moments. Turnover and average transaction costs are first calculated at the bond level and later summarized using medians. In turn, the slopes of transaction costs over trade size are computed using a unique regression for each mechanism subsample. To remove outliers, the sample of bonds is restricted to those that have at least ten observations for each moment computed. The final sample consists

[24] I also remove bonds that are security backed, equity-linked, putable, foreign-currency denominated, privately placed, perpetual, sold as part of a unit deal, or secured lease obligations bonds.

[25] Using standard letter-number equivalences (e.g., AAA=1, D=25), I average the letter ratings of the three agencies present in FISD: S&P, Moodie's and Fitch. I then go back to letter ratings using the same equivalence and classify as high yield a bond with a rating equal to or lower than BB+.

[26] Consider for example a dealer that performed four trades in a day, all of them with customers. In trade A the dealer sells 7K at 10:03 am, in trade B she buys 10K at 10:05 am, in trade C she sells 6K at 10:10 am, and in trade D she sells 3K at 10:10 am. In this case, the trades A, C, and D are competing to match with trade B. First I match by time distance, thus trades A and B form a pair. Trade A offsets all of its volume, so it is considered an agency trade. Trade B offsets 70% of its volume. The remaining 30% of the 10k are left to be matched with trades C and D. Given that these last trades happened at the same time, I match according to volume difference. Hence I form a pair with the remaining 3K of trade B and trade D. Again, trade D offsets all of its volume, so it is labeled as agency. Trade B offsets all of it volume as well, against A and D, so it is labeled as agency as well. In turn, trade C is labeled as principal.

of 2829 securities, which add up 1,602,438 observations. Subscripts t, b, d account for customer-dealer trades of a particular bond and during a specific day, respectively.

- Bond b monthly turnover:

$$\mathcal{T}_b = 100 \times \frac{\sum_t vol_{t,b}/iao_b}{k_b/30.5}, \tag{M.6}$$

where $vol_{t,b}$ is the notional volume in trade t, k_b is the day count after offering and before maturity within the period sample, and iao_b is the average amount outstanding during those k_b days. Note that this specification accounts for months in which the bond has no trades at all.

- Bond b volume-weighted average transaction cost in each mechanism:

$$\mathcal{S}_b^P = \sum_{t,d}(s_{t,b,d} \times vol_{t,b,d}^P)/\sum_{t,d} vol_{t,b,d}^P, \tag{M.7}$$

$$\mathcal{S}_b^A = \sum_{t,d}(s_{t,b,d} \times vol_{t,b,d}^A)/\sum_{t,d} vol_{t,b,d}^A, \tag{M.8}$$

where $s_{t,b,d}$ is Choi, Huh, and Shin (2023)'s Spread1:

$$s_{t,b,d} = Q \times 10000 \times (\frac{p_{t,b,d} - p_{b,d}^{DD}}{p_{b,d}^{DD}}) \quad , \quad p_{b,d}^{DD} = \frac{\sum_{t \in DD_{b,d}} vol_{b,d,t}^{DD} p_{b,d,t}^{DD}}{\sum_{t \in DD_{b,d}} vol_{b,d,t}^{DD}}$$

with $Q = 1$ (-1) if a customer buys (sells). To reduce the noise coming from micro trades, I only consider trades in which the volume >\$100K (Pinter, Wang, and Zou, 2022). Since prices are expressed per fixed amount of bond units, the percentage difference between the customer-dealer price $p_{t,b,d}$ and the inter-dealer price $p_{b,d}^{DD}$ equals the transaction costs per dollar computed in the model.

- Transaction cost - trade size slope in each mechanism. I estimate the following model for each mechanism subsample:

$$s_{t,d,b} = \alpha + \beta FE + \gamma 100(vol_{t,b,d}/iao_b) + \epsilon_{t,b,d}, \tag{M.9}$$

where $FE = [dealer, bond, day]$. Given that in the model the asset supply is normalized, to match the theoretical counterpart I consider the ratio between the volume traded and the amount outstanding. In that regard, the OLS estimates $\hat{\gamma}^P$ and $\hat{\gamma}^A$ are interpreted as how many bps transaction costs increase with a one percentage point increase in the traded amount outstanding of the bond. Appendix B.2 presents the regression results.

5.3 Estimation results

Table 2 presents the estimation results. To the best of my knowledge, this is the first paper to structurally estimate a search model using the US corporate bond secondary market data. Given this lack of reference, I limit the exposition to explain what the parameter values mean for our model and, when possible, trace comparisons with empirical observations.

Table 2: Baseline Calibration. Unit of time = 1 month

Parameter	Description	Value
- Normalized-		
A	Asset supply	1
ϵ_i	Preference shifter	$\left\{\frac{i-1}{I-1}\right\}_{i=1}^{20}$
- Externally calibrated-		
r	Discount rate	0.5%
π_i	Preference shifter distribution	$1/I$
η	Dealer's bargaining power	0.95
- GMM calibrated-		
α	Contact with dealer rate	9.15
δ	Preference shock rate	2.59
β	Agency execution rate	1.00
θ	Inventory cost (bp)	0.89
σ	Utility curvature	2.73

The estimation results tell us that search frictions matter. This is not only because customers need to wait a significant amount of time to contact dealers, but also because when they do so, they only partially realize their gains from trade. Customers contact dealers around 9 times per month, which means that they have to wait around 2 business days for an opportunity to update their holdings. In the model, the rate at which customers contact dealers is as important as the trade surplus they can preserve after paying transaction costs. In other words, what matters is the bargaining-adjusted rate, $\alpha(1-\eta)$. This later rate tells us that customers need to wait around 2 months to fully extract all the trading surplus from rebalancing positions.

Preference shocks happen with less intensity than trade opportunities. On expectation, a customer changes preferences around 2.5 times per month[27]. On the one hand, the fact that customers change preferences less often than the rate at which they contact dealers means that not all trading opportunities are realized. On the other hand, whenever trading does happen, the amount exchanged is larger than what it would be with a higher preference shock rate: customers can take more extreme positions knowing that those

[27] While preference shocks arrive at a Poisson rate of 2.59, the probability of receiving a preference different from the current one is 95%, given the uniform distribution and a support of 20 types.

positions will stay optimal longer. In turn, these larger amounts exchanged translate into higher transaction costs. Both infrequent trading and high transaction costs are salient features of the secondary corporate bond market.

In comparison with LR09, two parameters are added. The first of them is β, which accounts for the expected execution delay of an agency trade. The available data inform us only of when trades are executed, but not on the initial customer-dealer contact that started the transaction. The estimation results shed light on this scarcely explored parameter and tell us that the execution waiting times are considerable. Dealers take on expectation one month to execute trades for those customers not willing to pay the principal premium.

Regarding the second novel parameter, the marginal inventory costs θ, the results suggest that these are considerable: 0.89 bps for a one-way trade. To interpret this number, let me focus on the regulation-induced costs dealers face when including assets in their inventories. The empirical evidence indicates that the leverage ratio requirement (LRR) is the most tightly binding constraint for most U.S. banks after the post-2008 financial crisis regulations were set (Duffie, 2017; Greenwood, Stein, Hanson, and Sunderam, 2017). The LRR requires banks to hold capital for an amount of 5% of the non-risk-weighted value of assets in inventory[28]. Restricting attention to this most binding regulation, the inventory cost faced by a dealer buying $p(a' - a)$ worth of assets, with an average holding period of 10.6 days (Goldstein and Hotchkiss, 2020) and incurring a daily opportunity cost of $r/30\%$, is $5\%[p(a' - a)(e^{(r/30)10.6} - 1)]$. The model counterpart of such round-trip principal trade cost would be $2\theta_{LRR}p(a' - a)$, where θ_{LRR} would consider only this specific but important piece of regulation. The following mapping is obtained: $\theta_{LRR} = 5\%[e^{(r/30)10.6} - 1]/2 = 0.44$bps. The comparison between the estimated marginal inventory costs and this back-of-the-envelope LRR cost indicates that the estimation is in the right order of magnitude, arguably capturing other non-regulatory inventory costs.

Finally, the curvature of the utility function is estimated at 2.73. This parameter is related both to the intensive margin of trading and to the marginal trading surplus. As preferences approach the linear case, the amounts traded increase, with low (high) preference customers selling (buying) as much as possible. On the other hand, as preferences become linear, the marginal surplus from trading an extra unit becomes constant. The estimated value suggests that when customers rebalance positions, they do so in a moderate way, and that the marginal surplus from trading is increasing.

Table 3 presents the comparison between the theoretical moments and the empirical ones. Although existing tensions in the model prevent it from perfectly matching the targeted parameters, the results tell us that the model can fairly represent the stylized facts this paper is interested in.

[28]The percentage is 3% for non-global systemically important banks with assets over 250 billion dollars, and 5% for global systemically important banks.

Table 3: Model Fit

Moment	Empirical			Theoretical
	p50 (m_s)	p25	p75	$m(\hat{v})$
$\mathcal{S}^P$, Principal Vol Weighted Avg Costs	9.12	5.87	14.20	10.29
$\mathcal{S}^A$, Agency Vol Weighted Avg Costs	5.00	2.56	8.73	4.04
$\mathcal{T}$, Monthly Turnover	3.27	2.28	4.61	3.47
	$\hat{\gamma}$ (m_s)	$\hat{\gamma} - s.e.$	$\hat{\gamma} + s.e.$	
γ^P, Principal Cost-Size slope	1.45	1.33	1.58	1.31
γ^A, Agency Cost-Size slope	0.61	0.50	0.73	0.69

Note: Theoretical moments are computed at the steady state, using the calibration presented in Table 2. Empirical volume-weighted average cost and monthly turnover are computed at the bond level and summarized by computing the median and interquartile range. Empirical transaction costs - trade size slope is computed estimating equation (M.9)

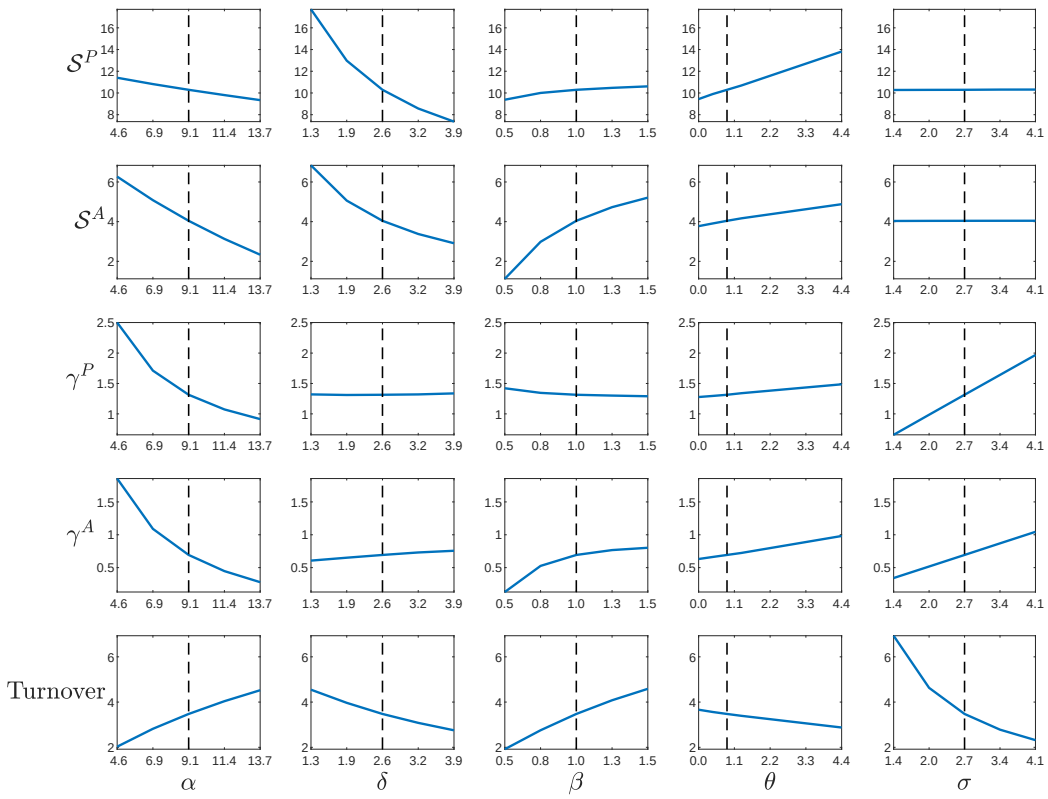
5.4 Identification

In this subsection I argue that the moments chosen are informative to jointly pin down the parameter values. In this regard, a common feature in search models of financial markets is the prevalence of general equilibrium effects. Typically, assets are valued according to the utility flow and trading opportunities they generate while customers travel across the state space. Particularly for the model here presented, an asset position would also determine the likelihood of choosing a given trading mechanism. This model structure implies that all parameters affect directly or indirectly the policy functions and correspondingly the observable moments generated. Figure 5 shows that, despite these general equilibrium effects, the different directions and intensities in which parameters and moments relate allow me to draw a unique mapping between them.

The first column of Figure 5 tells us how the theoretical moments change as we shift the contact rate with dealers. For this, I solve the model for alternative values of α while keeping all other parameters at their estimated values. Not surprisingly, turnover is increasing in the contact rate. The extensive margin increases as more contacts allow customers to trade more often. The intensive margin also increases as optimal asset holdings become more extreme: the expected time of holding unwanted positions is reduced. Perhaps less obvious is the diminishing effect α has on average transaction costs and on cost-size slopes. These figures decrease mainly for the same reason, the surplus from trading is reduced as trading opportunities become more frequent.

The rate at which customers receive preference shocks has the opposite effect on turnover. Although the extensive margin increases – the fraction of customers that contact dealers holding unwanted positions increases – the decrease in the intensive margin dominates. The latter is due to customers opting for less extreme positions, in anticipation of more frequent preference shocks. Regarding transaction costs, as customers expect to change preferences more often, the trading surplus decreases, and so transaction costs decrease as well. The relation between costs and trade size remains mostly unaffected by this parameter.

Figure 5: Theoretical Moments Variation.



Note: This figure depicts the theoretical moments' variation as parameters change around their estimated values, which are presented with vertical dashed lines. These parameters are the contact with dealers rate, α, the preference shock rate, δ, the agency execution rate, β, the inventory cost expressed in basis points, θ, and the utility curvature, σ. Unchanged parameters are set at their estimated values.

This last (lack of) effect hints at why including both average transaction costs and transaction costs - trade size slopes helps to identify the parameters. For example, changes in α affect trading size and marginal trading surplus / transaction costs in opposite directions, thus affecting the transaction costs - trade size slopes. Contrastingly, shifts in δ move both in the same direction, without significant effects on the implied slopes.

The third parameter in Figure 5 is the execution rate of agency trades, β. Similar to the effect of α on turnover, increasing the execution rate increases the extensive margin of both agency and principal turnover. Although optimal asset positions do not significantly change, migration across mechanisms and other general equilibrium effects (see section 6.2) imply that the intensive margin also increases. Consequently, the overall effect on turnover is positive. As expected, a change in the execution rate does not have a major impact on principal transaction costs or on principal cost-size slope. An increase in β makes the agency contract more valuable, given that customers will hold unwanted positions for less time, so both agency average transaction

costs and transaction costs derivative on trade size increase. The contrasting effect that the execution rate has on agency and principal related moments is the main source of identification of this parameter.

In turn, an increase in marginal inventory costs θ decreases turnover and increases principal transaction costs, in line with the empirical literature findings (see subsection 1.1). Due to general equilibrium effects, agency transaction costs increase as well. Basically, a less dispersed equilibrium asset distribution makes the waiting stage for agency trades less costly. Agency trading surpluses increase and dealers bargain larger transaction costs. This will be explained in detail in section 6.1.

Finally, the curvature of the utility function σ, as previously anticipated, plays two main roles. Firstly, as preferences become linear the optimal asset positions become more dispersed and the average trade size becomes larger. This effect increases turnover. Secondly, a lower curvature is translated into lower marginal trade surpluses and hence into lower marginal transaction costs. Therefore costs-size slopes decrease. What distinguishes σ from other estimated parameters, and hence accounts for its main source of identification, is the fact that this parameter does not affect average transaction costs. On the one hand, customers trade larger amounts thus they pay larger transaction costs. On the other hand, conditional on trade size, transaction costs decrease. Overall, these two effects cancel out, resulting in a null effect over average transaction costs.

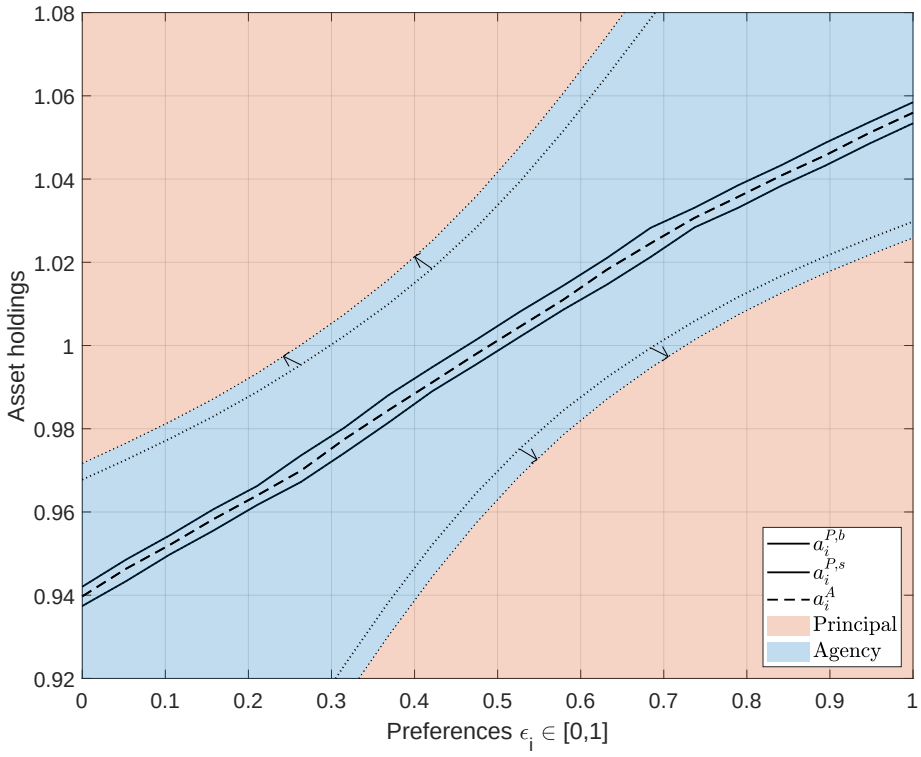
6 Numerical Exercises

In this section, I use the estimated model to revisit the evidence related to the two major changes observed in the US corporate bond markets in the last decade. First, I address the introduction of post 2008 financial crisis regulations by increasing the models' inventory costs. Second, motivated by the rising popularity of electronic trading venues, I analyze the effects of reducing the execution delay of agency trades. In both cases, when the economy moves through the parametric space, migration across mechanisms appears. Using the proposed decomposition, I show that composition effects account for an economically significant fraction of the changing costs.

6.1 Increase in Inventory Costs

An extensive empirical literature has shown how the stricter regulations implemented in the aftermath of the 2008 financial crisis increased dealers' inventory costs, raised the cost of principal trades and shifted volume towards larger agency intermediation (Anderson and Stulz, 2017; Schultz, 2017; Bao, O'Hara, and Zhou, 2018; Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; Dick-Nielsen and Rossi, 2019; Choi, Huh, and Shin, 2023). Here I revisit such evidence using the tools previously developed. I initially set the inventory costs to a smaller value, $\theta = 0.1$ bps, and then I increase it towards the estimated one. Figure 6

Figure 6: Policy functions as inventory costs increase.



Note: This figure depicts the policy functions of each customer, conditional on her preference type and current holdings, considering $\theta = 0.89$ bps. The lower and upper solid lines represent the buyer's and seller's optimal asset holdings under the principal trade, $a^{P,b}$ and $a^{P,s}$, respectively. The dashed line represents the optimal asset holdings under the agency trade, a^A. Regarding the mechanism choice, the principal and agency regions are shaded in orange and blue, respectively. To ease the comparison across calibrations, the trading mechanism thresholds under $\theta = 0.1$ bps, are depicted as dotted lines within the agency region, and the arrows denote its expansion.

shows the policy functions change as we increase inventory costs.

An increase in dealers' inventory costs makes principal trades more expensive. As a consequence, customers migrate towards agency trading. To highlight such migration, Figure 6 includes the low inventory cost case thresholds as dotted lines within the baseline calibration agency region. As can be seen, the agency region expands, being the migrating customers those with smaller trading needs [29].

Figure 7 presents the liquidity measures computed for $\theta \in [0.1\text{bps}, 0.89\text{bps}]$. Panel A shows that, as inventory costs increase, the overall turnover (black solid line) decreases. This is due to the combination of both extensive and intensive margins going in the same direction. On the one hand, fewer principal trades are being performed, due to the migration towards agency. Given the delayed execution of agency trades, overall daily trading decreases. On the other hand, the larger effective prices of principal trading make

[29]The optimal asset positions are also affected by an inventory costs increase. Such change is depicted in Appendix C.1. Since principal trading becomes more expensive, the trade size decreases: buyers (sellers) have lower (higher) optimal asset positions.

the average volume per trade decrease in such a mechanism and in the entire distribution. As expected, a positive relation between inventory costs and agency share (blue solid line) is present, which is explained by the aforementioned migration of trades.

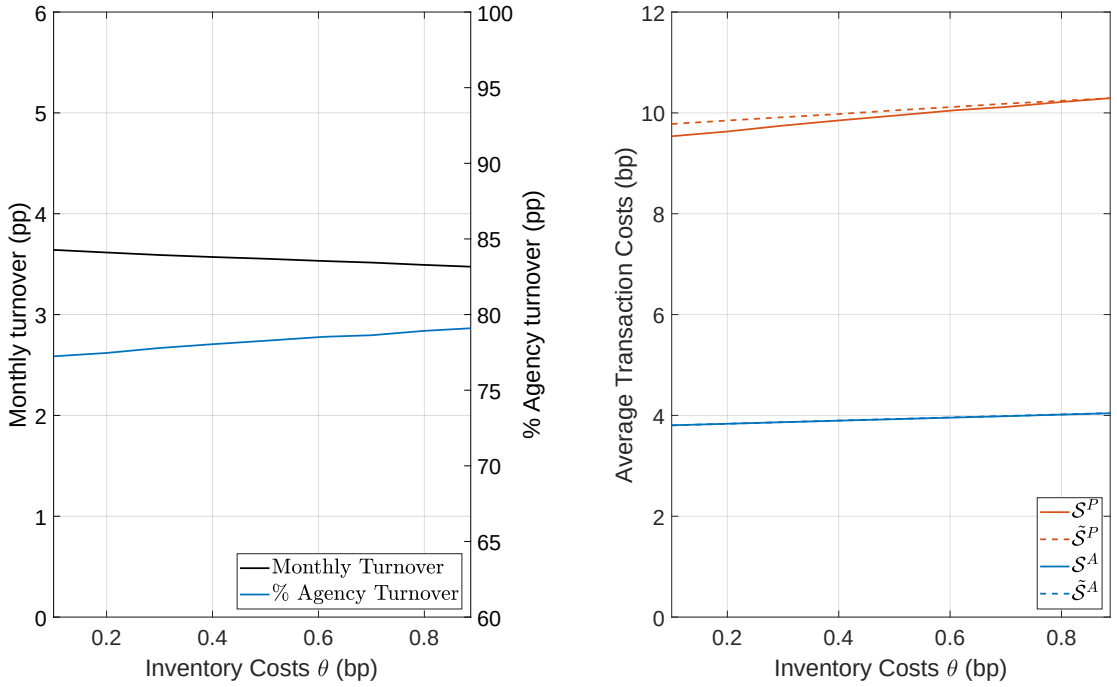
Transaction costs are jointly determined with trading volumes. Panel B presents the average costs for each mechanism, $\mathcal{S}^P$ and $\mathcal{S}^A$, in solid lines. As inventory costs rise, dealers translate a fraction of such increase through higher transaction costs, and so principal trading costs mechanically rise. Comparing the two extremes of the inventory costs range considered, the average principal cost increases by $\Delta\mathcal{S}^P = 0.76$ bps. Even though agency trades are not directly related to inventory costs, the transaction cost of these trades increase as well, by 0.24 bps. As was previously explained, the effect of inventory costs on agency costs is due to a general equilibrium effect. Given that fast trading becomes more costly, customers expect to hold their positions for longer. Therefore, when choosing these positions, they do so more moderately and the asset dispersion shrinks (see figure C.1). This implies that the burden of holding unwanted positions during the agency trade decreases, increasing both the agency surplus and its transaction costs.

The correlations between inventory costs, migration across mechanisms and average transaction costs have been broadly documented by both the empirical and the theoretical literature. Contrastingly, the self-selection of such migration and the consequent composition effect on cost measures has been largely overlooked. Panel B of Figure 7 accounts for such composition effects using the proposed decomposition. I use dashed lines to plot the counterfactual composition-free measures, $\tilde{\mathcal{S}}^P$ and $\tilde{\mathcal{S}}^A$, for each trading mechanism. The comparison of average and counterfactual measures allows us to gauge the sign and size of the bias.

Let me start by addressing principal costs. The migration pattern presented in Figure 6 tells us that principal customers can be split into non-migrants and outflowing migrants. When marginal inventory costs are set at $\theta = 0.1$ bps, the composition-free measure, i.e., the transaction cost paid by non-migrants, is already 0.24 bps larger than the mechanism's average. Such difference is understood going back to Figure 6, where it is observed that non-migrant principals are customers with relatively more extreme preferences and more extreme asset positions, both characteristics associated with higher transaction cost payments. As inventory costs increase, some customers migrate towards agency trading and the proportion of non-migrants increases. This process happens until the entire principal sample is composed by non-migrants. Mechanically, at the highest inventory cost considered, the composition-free and the average measures are equal. Therefore, the change in composition-free transaction costs is smaller than that of the mechanism's average: $\Delta\tilde{\mathcal{S}}^P = 0.51$bp. The difference is explained by a composition effect of $CE^P = 32.2\%$ [30]. In other words, when inventory costs increase, the average willingness to pay of the resulting sample increases, given that those customers who remain trading on principal are the ones who had a higher willingness to pay before costs increased. Therefore, the average transaction cost change captures this increase in the average

[30] Whenever $\tilde{\mathcal{S}}^P$ and $\mathcal{S}^P$ are linear on θ, the composition effect bias is constant. Figure 7 indicates that the computed slopes can be well approximated by linear functions.

Figure 7: Liquidity measures as inventory costs increase.



Note: Panel A (left) presents the steady-state total daily turnover rate, $\mathcal{T}$, and the agency percentage of such figure, $\mathcal{T}^{\mathcal{A}}/\mathcal{T}$, across $\theta \in [0.1\text{bps}, 0.89\text{bps}]$. Panel B (right) presents the steady-state volume-weighted average transaction costs for both mechanisms across $\theta \in [0.1\text{bps}, 0.89\text{bps}]$. Solid lines represent the average measures, $\mathcal{S}^{\mathcal{P}}$ and $\mathcal{S}^{\mathcal{A}}$, whereas dashed lines represent the counterfactual composition-free measures, $\tilde{\mathcal{S}}^{\mathcal{P}}$ and $\tilde{\mathcal{S}}^{\mathcal{A}}$.

willingness to pay, and is consequently biased upwards.

Regarding agency trades, the migration pattern associated with increasing inventory costs tells us that customers in this mechanism can be separated into non-migrants and inflowing migrants. At $\theta = 0.1$ bps, the entire agency sample is composed by non-migrants. Therefore, at such parametrizations composition-free and average costs are equal. As inventory costs increase, principal traders migrate towards agency, building up the proportion of inflowing migrants within the agency sample. At the highest inventory costs considered, I find that agency non-migrants pay only 0.07% higher costs than the mechanisms' average. This mild difference contrasts with the principal case, and it is explained by the small transaction costs dispersion found within agency customers, which implies that inflowing migrants pay similar costs to non-migrant customers (see Figure A.1). Given this similarity, composition effects are not expected to play an important role in agency transaction cost measures. As a matter of fact, when comparing the two extremes of the parametric range considered, the composition-free measure equals $\Delta\tilde{\mathcal{S}}^A = 0.242$bp, only 0.003bp above $\Delta\mathcal{S}^A$. Correspondingly, for the agency case, I find a mild composition effect bias of $CE^A = -1.2\%$.

To sum up, the model's predictions are in line with both the empirical and the theoretical literature that studies the effects of raising the intermediaries' inventory costs. In a nutshell, the provision of inventory-

related services becomes more expensive, and intermediation shifts away from principal towards agency trading. Nevertheless, the exercise also shows that transaction cost measures should be revisited, considering the impact that composition effects may have on them. Specifically, I find that these effects account for around a third of the increase in principal costs, and for a negligible figure on agency cost increases.

6.2 Decrease in the Execution Delay

Corporate bonds have been traditionally traded via voice messages. However, electronic platforms in which customers and dealers can contact counterparties simultaneously have gained popularity in the last decade. Not surprisingly, the empirical research shows that the increase in electronic trading made it easier for dealers to match counterparties in agency trades. Not only the agency share is higher for those bonds that are traded electronically, but also dealers use electronic platforms to find suitable counterparties for customers that contacted them through traditional voice messages (O'Hara and Zhou, 2021). From a customer's perspective, the increasing electrification of the market implies that dealers can find a matching trading counterparty faster. Thus, I model this market innovation as a reduction in the execution delay of such mechanism [31]. Such delay is captured in the model by β. In the estimated calibration, customers wait on expectation one month to execute their trades. I use the model to analyze the impact of decreasing three times such delay. The new policy functions are presented in Figure 8.

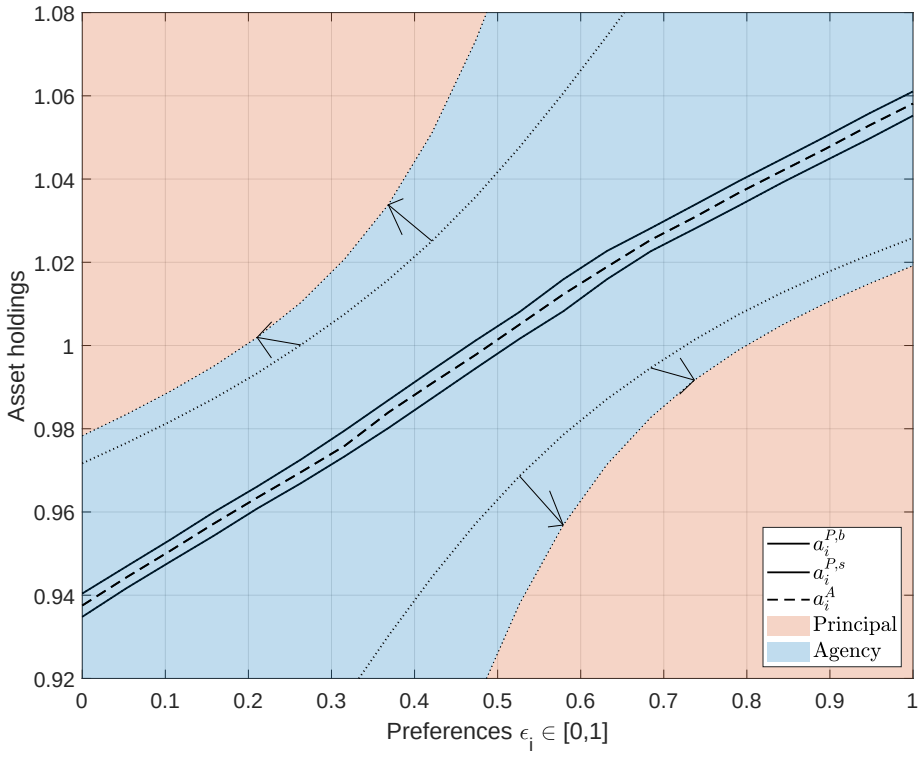
Figure 8 shows that a reduction in the waiting for execution time affects the trading mechanism choice. Smaller execution delays imply that agency customers need to hold unwanted positions for less time, thus the relative attractiveness of such a contract increases. Consequently, customers with preference type - asset positions close to the baseline calibration thresholds migrate away from principal towards agency.

The liquidity measures computed for the range $\beta \in [1, 3]$ are presented in Figure 9. Panel A presents the daily turnover as well as the percentage explained by agency trades. Increasing the execution speed of non-immediate contracts largely affects the extensive margin of both principal and agency trading. On the one hand, the number of customers that signed an agency contract can trade faster. On the other hand, the mass of customers waiting for execution is reduced; therefore, more customers are able to contact dealers in any given month and optimally choose whether to arrange new principal or new agency contracts. A less obvious effect of reducing execution delays is the decrease in the intensive margin of agency trading compared to that of principal. Firstly, the migrating customers make the average volume traded in both principal and agency contracts larger. Figure 8 shows, for each preference type, the expansion of both the maximum and the minimum trading size under agency and principal trades, respectively[32]. Secondly, a faster execution implies that agency customers are more likely to avoid a preference shock while waiting for

[31] Note that an alternative and non-mutually exclusive interpretation is a reduction in dealers' searching and matching costs, which are absent in my model.

[32] The optimal asset positions in the baseline and in the new calibration do not depart significantly, given that optimal assets are decided at execution in both mechanisms.

Figure 8: Policy function as execution delays decrease



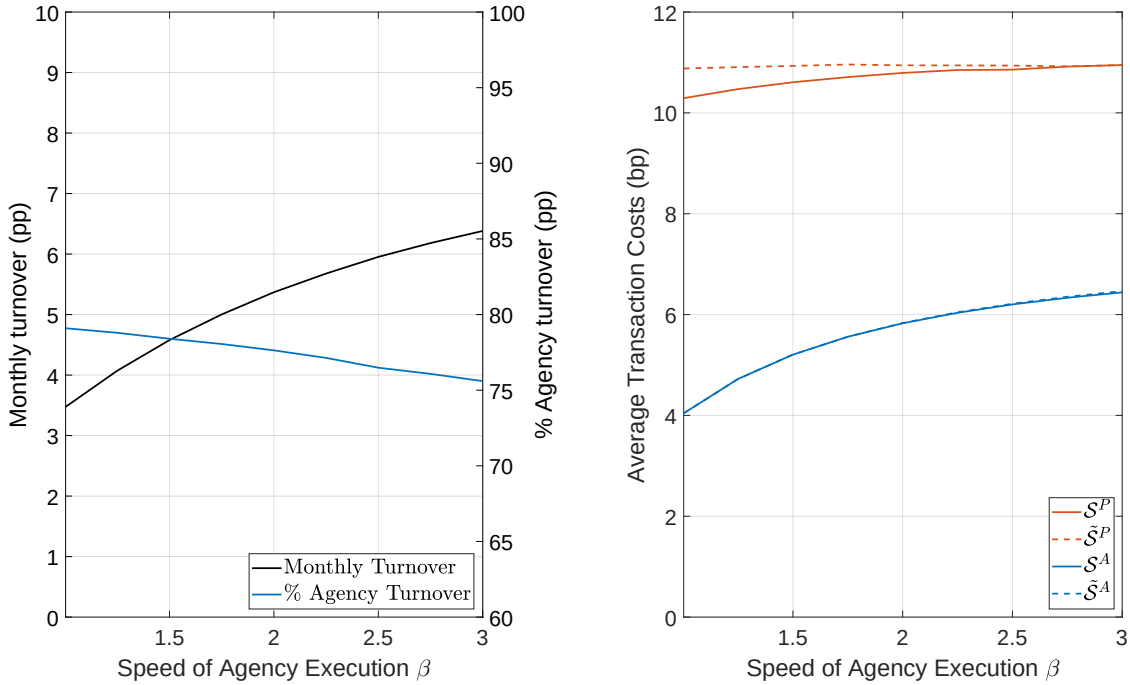
Note: This figure depicts the policy functions of each customer, conditional on her preference type and current holdings, considering $\beta = 3$. The lower and upper solid lines represent the buyer's and seller's optimal asset holdings under the principal trade, $a^{P,b}$ and $a^{P,s}$, respectively. The dashed line represents the optimal asset holdings under the agency trade, a^A. Regarding the mechanism choice, the principal and agency regions are shaded in orange and blue, respectively. To ease the comparison across calibrations, the trading mechanism thresholds under $\beta = 1$ are depicted as dotted lines within the agency region, and the arrows denote its expansion.

execution and trade according to their current preference types. Given that, in the steady state, the majority of the population is concentrated at the optimal asset positions, more customers trading according to the current type implies a decrease in the average agency volume per trade [33]. Overall, these effects jointly explain an increase in the daily turnover and a decrease in the agency share.

Panel B of Figure 9 shows the transaction costs in both mechanisms. Again, I decompose these figures into average and composition-free measures, which are depicted in solid and dashed lines, respectively. As execution delays decrease, average costs in both mechanisms go up. Principal costs increase by $\Delta\mathcal{S}^P = 0.66$ bps and agency costs rise by $\Delta\mathcal{S}^A = 2.4$ bps. Although speeding up agency trades makes trading in both mechanisms more expensive, the causes behind each of these changes are different. Regarding principal trades, the new calibration considered has no significant impact on the implied trading surplus of each

[33] LR09 contains a similar channel by which an increase in the contact rate with dealers, α, produces a steady state with a bigger accumulation of customers at their optimal positions, decreasing thus the average volume per trade.

Figure 9: Liquidity measures when execution delays decrease.



Note: Panel A (left) presents the steady-state total daily turnover rate, $\mathcal{T}$, and the agency percentage of such figure, $\mathcal{T}^{\mathcal{A}}/\mathcal{T}$, across $\beta \in [1,3]$. Panel B (right) presents the steady-state volume-weighted average transaction costs for both mechanisms across $\beta \in [1,3]$. Solid lines represent the average measures, $\mathcal{S}^{\mathcal{P}}$ and $\mathcal{S}^{\mathcal{A}}$, whereas dashed lines represent the counterfactual composition-free measures, $\tilde{\mathcal{S}}^{\mathcal{P}}$ and $\tilde{\mathcal{S}}^{\mathcal{A}}$.

customer. Therefore, keeping samples constant, principal costs should not significantly change. Accordingly, the counterfactual composition-free measure of principal costs has only a slight increase of $\Delta\tilde{\mathcal{S}}^P = 0.07$bp and almost the entire increase in average principal costs is due to composition effects, $CE^P = 89.54\%$. The explanation is found in Figure 8. Principal customers with relatively moderate preferences and asset positions, characteristics associated with low transaction cost payments, migrate away from the mechanism, increasing the average willingness to pay of the remaining principal sample. Regarding agency trades, a reduction in expected delays has a direct positive impact on the expected trade surplus of every agency customer: unwanted positions can be exchanged faster. I compute an increase in the agency composition-free costs of $\Delta\tilde{\mathcal{S}}^A = 2.40$ bps. Note that this figure is slightly higher than the average measure, which indicates that inflowing migrating customers have a slightly smaller trade surplus than the non-migrant agency customers. The corresponding composition effect bias is negligible, computed at $CE^A = -1.03\%$.

The results here obtained provide new insights about the impact that electronic venues have in OTC markets. By reducing execution delays, these platforms produce a shift in the demand towards agency trades, thus raising the transaction costs of such a mechanism. An effect over principal costs is also observed, which operates exclusively through composition effects. As customers shift their demand towards agency, the sample

of principal traders is reduced and the average surplus from trading on such mechanism increases. Therefore average immediacy costs spuriously increase. Although not studied here, the demand shifts observed arguably complement movements in the relative supply of trading mechanisms, due to the decrease in search and matching costs faced by dealers.

7 Conclusion

OTC markets have undergone several changes during the last decade. Intermediation activities had been perturbed by both new regulations and new trading technologies, affecting the cost and the speed at which customers can trade. In this paper, I study how customers optimally face these changing conditions and the consequences of such reaction over market liquidity and its measurement.

I develop a quantitative search model in which I can explicitly study the customers' trading mechanism choices. I show that the speed-cost trade-off faced when choosing between principal and agency trades is solved based on customers' trading needs, and that such trading needs are translated to transaction cost measures. The fact that trading mechanisms and transaction costs are jointly determined presents an empirical challenge. Whenever market conditions change, customers endogenously migrate across mechanisms, thus altering the composition of the samples in which liquidity measures are computed.

To overcome such challenge, I build counterfactual liquidity measures in which composition effects are controlled for. I estimate the model using corporate bond transaction data and perform numerical exercises motivated by recent developments in that market. In those exercises, a fraction of principal customers migrate towards agency trading. Given that those principal customers who did not migrate paid on average higher transaction costs, the change in principal average costs is upward biased. In particular, composition effects account for a third of the change in principal transaction costs after an inventory costs increase, and for almost all of the change after an increase in execution speed. In turn, agency costs are barely affected by composition effects.

The results here obtained contribute to the debate of whether stricter financial regulations set after 2008 were welfare-improving. If the cost of immediacy has not increased as much as was previously thought, new regulations may have improved financial soundness at a lower expense.

References

Adrian, T., Boyarchenko, N., and Shachar, O. (2017). Dealer balance sheets and bond liquidity provision. *Journal of Monetary Economics*, *89*, 92–109.

An, Y. (2022). Competing with inventory in dealership markets. *Available at SSRN 3284836*.

An, Y., and Zheng, Z. (2023). Immediacy provision and matchmaking. *Management Science*, *69*(2), 1245–1263.

Anderson, M., and Stulz, R. M. (2017). Is post-crisis bond liquidity lower? Tech. rep., National Bureau of Economic Research.

Bao, J., O'Hara, M., and Zhou, X. A. (2018). The volcker rule and corporate bond market making in times of stress. *Journal of Financial Economics*, *130*(1), 95–113.

Bessembinder, H., Jacobsen, S., Maxwell, W., and Venkataraman, K. (2018). Capital commitment and illiquidity in corporate bonds. *The Journal of Finance*, *73*(4), 1615–1661.

Choi, J., Huh, Y., and Shin, S. S. (2023). Customer liquidity provision: Implications for corporate bond transaction costs. *Management Science*, *(forthcoming)*.

Cimon, D., and Garriott, C. (2019). Banking regulation and market making. *Journal of Banking & Finance*, *109*, 105653.

Coen, J., and Coen, P. (2022). A structural model of liquidity in over-the-counter markets.

Cohen, A., Kargar, M., Lester, B., and Weill, P.-O. (2022). Inventory, market making, and liquidity: Theory and application to the corporate bond market.

Dick-Nielsen, J., and Poulsen, T. K. (2019). How to clean academic trace data. *Available at SSRN 3456082*.

Dick-Nielsen, J., and Rossi, M. (2019). The cost of immediacy for corporate bonds. *The Review of Financial Studies*, *32*(1), 1–41.

Duffie, D. (2012). Market making under the proposed volcker rule. *Rock Center for Corporate Governance at Stanford University Working Paper*, (106).

Duffie, D. (2017). *Post-crisis bank regulations and financial market liquidity*. Banca d'Italia.

Duffie, D., Fleming, M. J., Keane, F. M., Nelson, C., Shachar, O., and Van Tassel, P. (2023). Dealer capacity and us treasury market functionality. *FRB of New York Staff Report*, (1070).

Duffie, D., Gârleanu, N., and Pedersen, L. H. (2005). Over-the-counter markets. *Econometrica*, *73*(6), 1815–1847.

Duffie, D., Gârleanu, N., and Pedersen, L. H. (2007). Valuation in over-the-counter markets. *The Review of Financial Studies*, *20*(6), 1865–1900.

Dyskant, L., Silva, A. F., and Sultanum, B. (2023). Dealer costs and customer choice. Mimeo.

Foucault, T., Pagano, M., and Röell, A. (2013). *Market liquidity: theory, evidence, and policy*. Oxford University Press, USA.

Friewald, N., and Nagler, F. (2019). Over-the-counter market frictions and yield spread changes. *The Journal of Finance*, *74*(6), 3217–3257.

Goldstein, M. A., and Hotchkiss, E. S. (2020). Providing liquidity in an illiquid market: Dealer behavior in us corporate bonds. *Journal of Financial Economics*, *135*(1), 16–40.

Greenwood, R., Stein, J. C., Hanson, S. G., and Sunderam, A. (2017). Strengthening and streamlining bank capital regulation. *Brookings Papers on Economic Activity*, *2017*(2), 479–565.

Hendershott, T., Li, D., Livdan, D., and Schürhoff, N. (2020). True cost of immediacy. *Swiss Finance Institute Research Paper*, (20-71).

Hugonnier, J. (2012). Speculative behavior in decentralized markets.

Hugonnier, J., Lester, B., and Weill, P.-O. (2020). Frictional intermediation in over-the-counter markets. *The Review of Economic Studies*, *87*(3), 1432–1469.

Kargar, M., Lester, B., Lindsay, D., Liu, S., Weill, P.-O., and Zúñiga, D. (2021). Corporate bond liquidity during the covid-19 crisis. *The Review of Financial Studies*, *34*(11), 5352–5401.

Kirkby, R. (2017). Convergence of discretized value function iteration. *Computational Economics*, *49*(1), 117–153.

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

Miao, J. (2006). A search model of centralized and decentralized trade. *Review of Economic dynamics*, *9*(1), 68–92.

O'Hara, M., and Zhou, X. A. (2021). The electronic evolution of corporate bond dealers. *Journal of Financial Economics*, *140*(2), 368–390.

Pinter, G., and Uslu, S. (2022). Comparing search and intermediation frictions across markets. *Johns Hopkins Carey Business School Research Paper*, (pp. 22–08).

Pinter, G., Wang, C., and Zou, J. (2022). Size discount and size penalty: Trading costs in bond markets.

Plante, S. (2021). Agency and principal trading in otc markets. *Mimeo*.

Rapp, A. C., and Waibel, M. (2023). Managing regulatory pressure: Bank regulation and its impact on corporate bond intermediation. *Available at SSRN 4500131*.

Saar, G., Sun, J., Yang, R., and Zhu, H. (2023). From market making to matchmaking: Does bank regulation harm market liquidity? *The Review of Financial Studies*, *36*(2), 678–732.

Schultz, P. (2017). Inventory management by corporate bond dealers. *Available at SSRN 2966919*.

Shen, J. (2015). *Exchange or OTC market: a search-based model of market fragmentation and liquidity*. Ph.D. thesis, Chapter 1, London School of Economics and Political Science.

Stokey, N. L., Lucas, R. E., and Prescott, E. C. (1989). *Recursive methods in economic dynamics*. Harvard University Press.

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

Weill, P.-O. (2020). The search theory of over-the-counter markets. *Annual Review of Economics*, *12*, 747–773.

Appendix A

A.1 Bargaining Outcomes

Here I compute the bargaining outcomes for the principal contract. The agency contract terms of trade can be obtained similarly.

$$\begin{aligned}[a_i^P(a), \phi_i^P(a)] &= \arg\max_{(a',\phi')} \Big\{ V_i(a') - V_i(a) - p(a'-a) - \phi' \Big\}^{1-\eta} \Big\{ \phi' - \theta p|a'-a| \Big\}^{\eta} \\ &= \arg\max_{(a',\phi')} (1-\eta) \ln[\underbrace{V_i(a') - V_i(a) - p(a'-a) - \phi'}_{A}] + \eta \ln[\underbrace{\phi' - \theta p|a'-a|}_{B}].\end{aligned}$$

$$\begin{aligned}\text{FOC}_{\phi'}: \qquad & -(1-\eta)A^{-1} + \eta B^{-1} = 0 \qquad \text{(assume interior solution)} \\ & \eta A - (1-\eta)B = 0 \\ & \eta[V_i(a') - V_i(a) - p(a'-a)] + (1-\eta)\theta p|a'-a| = \phi_i^P(a)\end{aligned}$$

Second-order conditions can be checked trivially, therefore $\phi_i^P(a)$ is the unique global maximizer. Now let us introduce the solution for $\phi_i^P(a)$ in the maximization argument to obtain (4).

$$\begin{aligned}a_i^P(a) = \arg\max_{a'} & \Big\{ (1-\eta) \big[V_i(a') - V_i(a) - p(a'-a) - \theta p|a'-a| \big] \Big\}^{1-\eta} \\ & \Big\{ \eta \big[V_i(a') - V_i(a) - p(a'-a) - \theta p|a'-a| \big] \Big\}^{\eta} \\ \arg\max_{a'} \quad & V_i(a') - V_i(a) - p(a'-a) - \theta p|a'-a|.\end{aligned}$$

A.2 Customer's Value Function Using Bargain-adjusted Contact Rate.

Here I show that the customer's value function can be rewritten as if the contact rate with dealers was $(1-\eta)\alpha$ and the customer had full bargaining power. In other words, the utility flow of an investor trading at α rate with a dealer with η bargaining power is equal to that of an investor trading at a slower rate $(1-\eta)\alpha$ with a dealer with no bargaining power. Let's replace the optimal terms of trade from equations (3), (4), (5) and (6) in equation (1).

$$V_{i(t)}(a) = \mathbb{E}_{i(t)}\Big[\int_t^{T_\alpha} e^{-r[s-t]}u_{i(s)}(a)ds$$
$$+ e^{-r[T_\alpha - t]}\max\Big\{(1-\eta)\big[V_{i(T_\alpha)}(a^P_{i(T_\alpha)}) - p(a^P_{i(T_\alpha)} - a) - \theta p|a^P_{i(T_\alpha)} - a|\big] + \eta V_{i(T_\alpha)}(a),$$
$$(1-\eta)V^A_{i(T_\alpha)}\big(a, \phi^A_{i(T_\alpha)}(a) = 0\big) + \eta V_{i(T_\alpha)}(a)\Big\}\Big].$$

Define the time it takes for a customer to receive either the preference shock or the contact with dealers shock as τ_δ and τ_α, respectively. These are exponentially distributed with their corresponding parameters δ and α. In turn, define $\tau = \min\{\tau_\delta, \tau_\alpha\}$. Now consider the above Bellman equation over some small time horizon h, and let h go to zero:

$$\begin{aligned}
V_i(a) =& \frac{1}{1+rh}\Big[u_i(a)h + Pr[\tau = \tau_\alpha \le h]\Big[(1-\eta)\max\Big\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a|, V_i^A\big(a, \phi_i^A(a) = 0\big)\Big\} + \eta V_i(a)\Big] \\
&+ Pr[\tau = \tau_\delta \le h]\Big[\sum_j \pi_j V_j(a)\Big] + Pr[\tau > h]V_i(a)\Big] \\
=& \frac{1}{1+rh}\Big[u_i(a)h + \alpha h\Big[(1-\eta)\max\Big\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a|, V_i^A\big(a, \phi_i^A(a) = 0\big)\Big\} + \eta V_i(a)\Big] \\
&+ \delta h\Big[\sum_j \pi_j V_j(a)\Big] + (1 - \delta h - \alpha h)V_i(a)\Big] \\
=& \frac{1}{1+rh}\Big[u_i(a)h + \underbrace{\alpha(1-\eta)h}_{Pr[\tau' = \tau_\kappa \le h]}\Big[\max\Big\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a|, V_i^A\big(a, \phi_i^A(a) = 0\big)\Big\}\Big] \\
&+ \delta h\Big[\sum_j \pi_j V_j(a)\Big] + \underbrace{(1 - \delta h - \alpha(1-\eta)h)}_{Pr[\tau' > h]}V_i(a)\Big],
\end{aligned}$$

where $\tau' = \min\{\tau_\delta, \tau_\kappa\}$ and τ_κ is the bargaining-adjusted time it takes to contact a dealer, which is exponentially distributed with parameter $\kappa = \alpha(1-\eta)$. Therefore, the customer's problem is represented by a Bellman equation were the contact with a dealer happens with Poisson arrival rate $(1-\eta)\alpha$, but where the customers have full negotiation power, $\eta' = 0$.

A.3 Expectations Resolution in the Flow Bellman Equation.

I keep on using τ_δ and τ_κ as the time it takes for a customer to receive either the preference shock or the (effective) contact shock, respectively, and $\tau' = \min\{\tau_\delta, \tau_\kappa\}$. In turn, define τ_β as the time it takes for a customer to be matched with another customer after choosing the agency trade. Consider the equation

derived in Appendix A.2 over some small time horizon h, and let h go to zero [34].

$$V_i(a) = \frac{1}{1+rh}\Big[u_i(a)h + Pr[\tau' = \tau_\delta \leq h]\sum_j \pi_j V_j(a)$$
$$+ Pr[\tau' = \tau_\kappa \leq h]\max\Big\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a|, V_i^A(a)\Big\} + Pr[\tau' > h]V_i(a)\Big]$$
$$V_i(a) = \frac{1}{1+rh}\Big[u_i(a)h + \delta h\sum_j \pi_j V_j(a)$$
$$+ \kappa h\max\Big\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a|, V_i^A(a)\Big\} + \big(1-(\delta+\kappa)h\big)V_i(a)\Big]$$
$$V_i(a)[1 + rh] = u_i(a)h + \delta h\sum_j \pi_j[V_j(a) - V_i(a)]$$
$$+ \kappa h\max\Big\{V_i(a_i^P) - V_i(a) - p(a_i^P - a) - \theta p|a_i^P - a|, V_i^A(a) - V_i(a)\Big\} + \cancel{V_i(a)}$$

$$rV_i(a) = u_i(a) + \delta\sum_j \pi_j[V_j(a) - V_i(a)] + \kappa\max\Big\{V_i(a_i^P) - V_i(a) - p(a_i^P - a) - \theta p|a_i^P - a|, V_i^A(a) - V_i(a)\Big\},$$

where $V_i^A(a)$ is the maximum utility a customer expects to get when she chooses the agency trade. Similarly, I can define this latter function in terms of flow utility as:

$$rV_i^A(a) = u_i(a) + \delta\sum_j \pi_j[V_j^A(a) - V_i^A(a)] + \beta\big[V_i(a_i^A) - V_i^A(a) - p(a_i^A - a)\big],$$

where $1/\beta$ is the time a customer expects to wait until the dealer finds him a counterpart and a_i^A is the optimal agency asset position chosen at execution (see equation (6)). Note that, while waiting, the customer might change his preferences, which is reflected in the second term on the right-hand side of the above equation. The expression $V_i^A(a)$ can be further manipulated to be written as a function of $V_i(a)$. Let me first obtain the expression for $\sum_j \pi_j V_j^A(a)$:

$$(r+\delta+\beta)V_i^A(a) = u_i(a) + \delta\sum_j \pi_j V_j^A(a) + \beta\big[V_i(a_i^A) - p(a_i^A - a)\big]$$
$$(r+\cancel{\delta}+\beta)\sum_i \pi_i V_i^A(a) = \sum_i \pi_i u_i(a) + \cancel{\delta\sum_j \pi_j V_j^A(a)} + \beta\sum_i \pi_i\big[V_i(a_i^A) - p(a_i^A - a)\big]$$
$$\sum_j \pi_j V_j^A(a) = \frac{1}{r+\beta}\Big[\sum_j \pi_j u_j(a) + \beta\sum_j \pi_j\big[V_j(a_j^A) - p(a_j^A - a)\big]\Big].$$

[34] For ease of exposition I removed time subscripts.

Plugging this result into $V_i^A(a)$ equation:

$$(r+\delta+\beta)V_i^A(a) = u_i(a) + \frac{\delta}{r+\beta}\Big[\sum_j \pi_j u_j(a) + \beta\sum_j \pi_j\big[V_j(a_j^A) - p(a_j^A - a)\big]\Big] + \beta\big[V_i(a_i^A) - p(a_i^A - a)\big]$$

$$V_i^A(a) = \underbrace{\frac{1}{r+\beta}\frac{(r+\beta)u_i(a) + \delta\sum_j \pi_j u_j(a)}{r+\delta+\beta}}_{\bar{U}_i^\beta(a)} + \underbrace{\frac{\beta}{r+\beta}}_{\hat{\beta}}\Big[\underbrace{\frac{(r+\beta)V_i(a_i^A) + \delta\sum_j \pi_j V_j(a_j^A)}{r+\delta+\beta}}_{\bar{V}_i^A} - p\big[\underbrace{\frac{(r+\beta)a_i^A + \delta\sum_j \pi_j a_j^A}{r+\delta+\beta}}_{\bar{a}_i^A} - a\big]\Big]$$

$$V_i^A(a) = \bar{U}_i^\beta(a) + \hat{\beta}\big[\bar{V}_i^A - p(\bar{a}_i^A - a)\big]$$

Finally, I can include this result in the initial equation, rearrange and define terms in a similar way as was previously done. The flow Bellman equation of a customer of type i holding assets a waiting to contact a dealer in any given period is the following:

$$\begin{aligned} V_i(a) = \bar{U}_i^\kappa(a) + \hat{\kappa}\Big[&[1-\hat{\delta}]\max\Big\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a|, \bar{U}_i^\beta(a) + \hat{\beta}\big[\bar{V}_i^A - p(\bar{a}_i^A - a)\big]\Big\} \\ &+ \hat{\delta}\sum_j \pi_j \max\Big\{V_j(a_j^P) - p(a_j^P - a) - \theta p|a_j^P - a|, \bar{U}_j^\beta(a) + \hat{\beta}\big[\bar{V}_j^A - p(\bar{a}_j^A - a)\big]\Big\}\Big], \end{aligned}$$

where $\bar{U}_i^\kappa(a) = \Big[\frac{(r+\kappa)u_i(a) + \delta\sum_j \pi_j u_j(a)}{r+\delta+\kappa}\Big]\frac{1}{r+\kappa}$, $\hat{\kappa} = \frac{\kappa}{r+\kappa}$ and $\hat{\delta} = \frac{\delta}{r+\delta+\kappa}$.

A.4 Trading Mechanism Choice Sets

After subtracting the common term $V_i(a)$, the indifference condition writes:

$$V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a| = \bar{U}_i^\beta(a) + \hat{\beta}\big[\bar{V}_i^A - p(\bar{a}_i^A - a)\big]$$

Firstly, consider the indifference condition for the cases where agents change their positions should they trade under the principal mechanism. Conditional on increasing or reducing positions, and disregarding the current valuation $V_i(a)$, the gains from a principal trade increase at a constant rate in current asset holdings a. This is a direct consequence of modeling constant dealers' marginal costs and can be seen on the left-hand side of the indifference equation. On the other hand, in the agency mechanism, the customer keeps his current asset holdings until some counterparty is found. Given decreasing marginal instant utility, the flow utility she derives while waiting for execution, $\bar{U}_i^\beta(a)$, marginally decreases in current asset holdings a. After the waiting period is over, the customer will obtain a discounted gain from trade, which is also linear in a,

since optimal agency holdings are independent of current holdings (see equation (6)). Therefore, the total gains from a delayed intermediated trade are marginally decreasing in a. I will exploit these differences in the two types of trades to find the current asset holdings thresholds as the roots of the indifference condition. Let us rearrange the arguments of such an indifference equation:

$$\underbrace{V_i(a_i^P) - p(1+\psi\theta)a_i^P - \hat{\beta}(\bar{V}_i^A - p\bar{a}_i^A)}_{B_i} = \underbrace{\bar{U}_i^{\beta}(a)}_{C_i(a)} + \underbrace{pa\big(\hat{\beta} - (1+\psi\theta)\big)}_{D(a)},$$

where $\psi = 1$ $(= -1)$ if $a_i^P - a \geq 0$ (< 0). The left-hand side, B_i, is independent of current asset holdings a, while the two arguments on the right-hand side are not. Firstly, $C_i(a)$ is a twice continuously differentiable, strictly increasing, and strictly concave function that satisfies Inada conditions in current asset holdings a. Secondly, $D(a)$ is linear in a, and its sign depends on the difference between the expected present value of reselling the asset through agency and reselling the asset immediately plus the inventory cost discount. Given the assumption made about the marginal inventory costs, $\theta < \frac{r}{r+\beta}$, $D(a)$ is a strictly decreasing linear function on a, and the right-hand side is thus inverse U-shaped [35].

Let us consider now the indifference condition for the cases where customers would not trade if they were to opt for principal trading. A customer that does not trade derives utility by holding his current position until the next contact with a dealer. In turn, an agency trader adds up the utility of holding his current position until the execution of the trade, plus the gains from trade she gets without paying an immediacy premium. As before, I can rearrange this indifference condition to express its components according to their dependence on the current position.

$$\underbrace{-\hat{\beta}(\bar{V}_i^A - p\bar{a}_i^A)}_{B_i} = \underbrace{\bar{U}_i^{\beta}(a) - V_i(a)}_{C_i(a)} + \underbrace{pa\hat{\beta}}_{D(a)}.$$

The left-hand side, B_i, is still independent of current asset holdings a_i. Regarding the right-hand side, $D(a)$ is linear and strictly increasing in a. In turn, $C_i(a)$ subtracts from a strictly increasing and strictly concave function a function $V_i(a)$ that, at this point, is unknown. The shape of $C_i(a)$ determines the region under which customers decide not to trade at all. Given the unavailability of close form solutions for the value function, these regions are characterized numerically. Under all different plausible calibrations, the numerical solution of the model indicates that $C_i(a) + D_i(a)$ is U-shaped.

This analysis indicates that the optimal trading mechanism choice for each preference type can be characterized by partitions of the subsets $\Gamma_i = \{Buy_i, Sell_i, NoT_i\}$, which in turn defined the optimal trading direction for a customer trading on principal. Formally, define $\hat{a}_i^{h,\rho}$, with $h = \{1, 2\}$ and $\rho = \{b, s, nt\}$, as the

[35] The parameter values discussed in the calibration section indicate that $\theta < \frac{r}{r+\beta}$ is not a binding restriction for most plausible calibrations.

current asset holdings that make customers of type i indifferent between the principal or the agency trade, where h denotes the threshold number and ρ indicates if the threshold is computed for a potential principal buyer, seller or non trader. In turn, define the partitions Γ_i^P and Γ_i^A as the type specific subsets of asset holdings within which a customer of type i would trade on principal or through agency in the steady state, respectively, for a specific principal trade direction $\Gamma_i = \{Buy_i, Sell_i, NoT_i\}$. The indifference condition provides two possible scenarios for each principal trade direction:

$$
Buy_i \begin{cases} B_i \geq C_i(a) + D_i(a) & \forall a: \quad \Gamma_i^P = \Gamma_i. \\ B_i < C_i(a) + D_i(a) & \text{for some } a: \quad \Gamma_i^P = \Gamma_i \cap \{[-\infty, \hat{a}_i^{1,b}] \cup [\hat{a}_i^{2,b}, \infty)\}, \ \Gamma_i^A = \Gamma_i \setminus \Gamma_i^P. \end{cases}
$$

$$
Sell_i \begin{cases} B_i \geq C_i(a) + D_i(a) & \forall a: \quad \Gamma_i^P = \Gamma_i. \\ B_i < C_i(a) + D_i(a) & \text{for some } a: \quad \Gamma_i^P = \Gamma_i \cap \{[-\infty, \hat{a}_i^{1,s}] \cup [\hat{a}_i^{2,s}, \infty)\}, \ \Gamma_i^A = \Gamma_i \setminus \Gamma_i^P. \end{cases}
$$

$$
NoT_i \begin{cases} B_i < C_i(a) + D_i(a) & \forall a: \quad \Gamma_i^P = \emptyset. \\ B_i \geq C_i(a) + D_i(a) & \text{for some } a: \quad \Gamma_i^P = \Gamma_i \cap \{[\hat{a}_i^{1,nt}, \hat{a}_i^{2,nt}], \ \Gamma_i^A = \Gamma_i \setminus \Gamma_i^P. \end{cases}
$$

A.5 Existence and Uniqueness of the Value Function.

In order to prove the uniqueness of the value function $V_i(a)$, I need to show that the Bellman operator T, defined as the right-hand side of (7), is a contraction mapping that operates in a Banach space, i.e., a complete normed vector space. To show completeness, I can rely on Theorem 3.1 in Stokey, Lucas, and Prescott (1989) - SL89 -, which requires the functions mapped by T to be continuous and bounded. Define $S = R_+ \times \{1, .., I\}$, $C = \{g : S \to R \mid g(a, i) \text{ is continuous in } a \text{ and bounded above}\}$ and the metric space $(C, \|.\|)$, where $\|.\|$ denotes the *sup norm*. I want the right-hand side of equation (7) to belong to C. By assumption, the utility function $u_i(a)$ is continuous, property preserved by the linear combination $\bar{U}_i^{\kappa}(a)$. Secondly, each term on the two sides of the max operator is continuous as well. Given the existence of thresholds $\bar{a}_i$ that make customers of type i indifferent between the two types of trade, both sides of the max operator return the same value at those thresholds. Hence, the utility a customer gets when her asset holdings change and cross a threshold does not suffer a jump. Finally, the stock of assets in the economy is in fixed supply $A \in R_+$, thus individual holdings are bounded. Therefore, $T : C \to C$ and $(C, \|.\|)$ is a complete metric space[36].

[36]The trading mechanism choice produces kinks in the value function. At those points, the value function will not be differentiable. Theorem 3.2 in SL89 only requires continuity, and that is guaranteed by the indifference condition that originates

Our next step is to show that this operator is a contraction mapping. I will rely on Blackwell's sufficient conditions (Theorem 3.3, SL89). Therefore, I need to show that the operator satisfies the monotonicity and discounting properties.

Monotonicity: Take any pair $V^1, V^2 \in C$ such that $V^1(i,a) \leq V^2(i,a)$, for all {a,i}$\in S$. I need to show that $[TV^1](i,a) \leq [TV^2](i,a)$, for all {a,i}$\in S$. From equation (7), the outcome of the max operators (decision of trade type) will always be greater or equal under $V^2(i,a)$ than under $V^1(i,a)$, since the arguments under both principal trade or agency are strictly increasing in the value function considered. The first term in equation (7) does not depend on the value function, and the second term is a convex combination of these max operators (with weights $(1-\hat{\delta})$ and $\hat{\delta}$ respectively), so the weak inequality holds and monotonicity is achieved.

Discounting: I need to demonstrate that there exist some $\lambda \in (0,1)$ such that $[T(V+\epsilon)](i,a) \leq [TV](i,a) + \lambda\epsilon$ for all $V \in C$, {a,i}$\in S$ and $\epsilon \geq 0$. Consider $[T(V+\epsilon)](i,a)$:

$$[T(V+\boldsymbol{\epsilon})](i,a) =$$

$$= \bar{U}_i^{\kappa}(a) + \hat{\kappa}\Big[[1-\hat{\delta}]\max\Big\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a| + \boldsymbol{\epsilon}, \bar{U}_i^{\beta}(a) + \hat{\beta}\big[\bar{V}_i^A - p(\bar{a}_i^A - a)\big] + \hat{\beta}\boldsymbol{\epsilon}\Big\}$$
$$+ \hat{\delta}\sum_j \pi_j \max\Big\{V_j(a_j^P) - p(a_j^P - a) - \theta p|a_j^P - a| + \boldsymbol{\epsilon}, \bar{U}_j^{\beta}(a) + \hat{\beta}\big[\bar{V}_j^A - p(\bar{a}_j^A - a)\big] + \hat{\beta}\boldsymbol{\epsilon}\Big\}\Big]$$

$$= \bar{U}_i^{\kappa}(a) + \hat{\kappa}\Big[[1-\hat{\delta}]\max\Big\{V_i(a_i^P) - p(a_i^P - a) - \theta p|a_i^P - a|, \bar{U}_i^{\beta}(a) + \hat{\beta}\big[\bar{V}_i^A - p(\bar{a}_i^A - a)\big] - \mathbf{(1-\hat{\beta})\boldsymbol{\epsilon}}\Big\}$$
$$+ \hat{\delta}\sum_j \pi_j \max\Big\{V_j(a_j^P) - p(a_j^P - a) - \theta p|a_j^P - a|, \bar{U}_j^{\beta}(a) + \hat{\beta}\big[\bar{V}_j^A - p(\bar{a}_j^A - a)\big] - \mathbf{[1-\hat{\beta}]\boldsymbol{\epsilon}}\Big\}\Big] + \hat{\kappa}\boldsymbol{\epsilon}$$

$$\leq [T(V)](i,a) + \hat{\kappa}\boldsymbol{\epsilon}$$

where the last inequality comes from the fact that subtracting a scalar to a component of a max operator will yield a weakly smaller value. To gain intuition, consider the parametrization case such that all customers, i.e., any pair {a,i}, choose the principal trade. In that case, $[T(V+\epsilon)](i,a) \leq [TV](i,a) + \hat{\kappa}\epsilon$, where $\hat{\kappa} = \kappa/(r+\kappa) \in (0,1)$. Alternatively, consider the parametrization under which every customer chooses the agency trade. In such case, $[T(V+\epsilon)](i,a) \leq [TV](i,a) + \hat{\kappa}\hat{\beta}\epsilon$, where $\hat{\kappa}\hat{\beta} \in (0,1)$ as well. Any case in between will yield a discounting factor between these two bounds $[\hat{\kappa}\hat{\beta}, \hat{\kappa}]$.

the kinks. See Kirkby (2017) for a proof of the convergence of the computational solution to the true solution using discretized value function iteration.

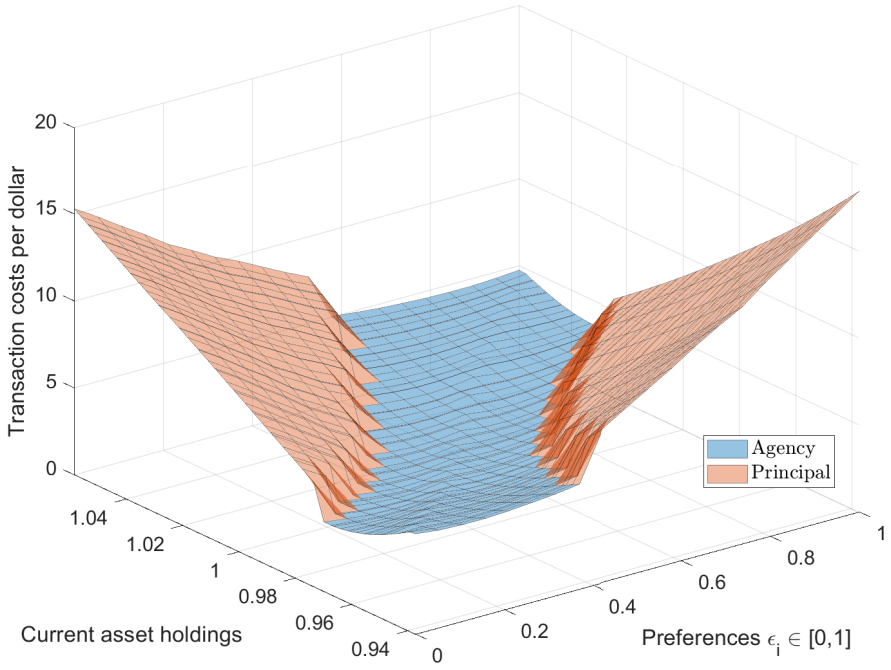
A.6 Solution Method Algorithm

The steady state of the model for any given inter-dealer price, p, is solved using the value function iteration method, enhanced with Howard's improvement step. The obtained policy and value functions, conditional on p, are nested within the computation of the market clear condition 14 to obtain the equilibrium inter-dealer price. The algorithm can be described by the following steps:

1. Set an initial guess for the equilibrium price p.
   (a) Set an asset holdings grid and an initial guess for $V_i(a)$
   (b) Compute optimal asset holdings $\{a_i^P(a), a_i^A\}_{i=1}^I$ using equations (4) and (6).
   (c) Compute trading mechanism choice for each pair $\{i, a\}$, using equation (8).
   (d) Fix $\{a_i^P(a), a_i^A\}_{i=1}^I$, and iterate h times the following steps:
      i. Update $V_i(a)$ using equation (7).
      ii. Compute trading mechanism choice for each pair $\{i, a\}$, using equation (8)
   (e) Update $V_i(a)$ using equation (7) until convergence.
2. Define trading mechanism sets $\{\Gamma_i^P, \Gamma_i^A\}_{i=1}^I$ using equation (8).
3. Compute transition matrix T using Equations (9), (10), (11), (12) and (13).
4. Set vector n_0 and obtain $n = \lim_{k \to K} n_0 T^k$, with K sufficiently large to reach convergence.
5. Compute aggregate gross demand and update p until excess demand in equation (14) converges towards zero.

A.7 Transaction Costs per Dollar Traded

Figure A.1: Transaction costs per dollar traded under each trading mechanism.



Note: This figure depicts the transaction costs per dollar traded paid by each customer, conditional on her preference type and current holdings, and expressed in basic points. Agency transaction costs are computed using the expected volume traded for each customer, as is explained in subsection 4.1, and expressed in present value at the moment of contact with the dealer.

A.8 Transaction Costs Decomposition

Here I present the algebra steps needed to decompose the transaction cost measures in equations (17) and (18). Specifically, I decompose the transaction cost measures computed under some parametrization $q = 0$, considering an alternative parametrization $q = 1$. The decomposition of transaction costs computed for a different parametrization and considering a different alternative parametrization follow the same steps.

$$
\begin{aligned}
\mathcal{S}^{P,0} &= \sum_{i\in\mathcal{I}}\sum_{a\in P_i^0}\frac{n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}\frac{\phi^{P,0}_{a,i}}{|a_i^{P,0}-a|p^0} \\
&= \underbrace{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap P_i^1}\frac{n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap P_i^1} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}\frac{\phi^{P,0}_{a,i}}{|a_i^{P,0}-a|p^0}}_{\mathcal{S}^{P,0}_{P^0,P^1}}\times\underbrace{\frac{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap P_i^1} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}}_{w^{P,0}_{P^0,P^1}} \\
&+ \underbrace{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap A_i^1}\frac{n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap A_i^1} n^0_{[a,i,\omega_1]}|a_i^{P}-a|}\frac{\phi^{P,0}_{a,i}}{|a_i^{P,0}-a|p^0}}_{\mathcal{S}^{P,0}_{P^0,A^1}}\times\underbrace{\frac{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap A_i^1} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}}_{w^{P,0}_{P^0,A^1}} \\
&+ \underbrace{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap NT_i^1}\frac{n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap NT_i^1} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}\frac{\phi^{P,0}_{a,i}}{|a_i^{P,0}-a|p^0}}_{\mathcal{S}^{P,0}_{P^0,NT^1}}\times\underbrace{\frac{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0\cap NT_i^1} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i^0} n^0_{[a,i,\omega_1]}|a_i^{P,0}-a|}}_{w^{P,0}_{P^0,NT^1}} \\
&= \mathcal{S}^{P,0}_{P^0,P^1}\times w^{P,0}_{P^0,P^1}+\mathcal{S}^{P,0}_{P^0,A^1}\times w^{P,0}_{P^0,A^1}+\mathcal{S}^{P,0}_{P^0,NT^1}\times w^{P,0}_{P^0,NT^1}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{S}^{A,0} &= \sum_{i\in\mathcal{I}}\sum_{a\in A_i^0}\frac{n^0_{[a,i,\omega_1]}rav^0_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0} n^0_{[a,i,\omega_1]}rav^0_{a,i}}\frac{\phi^{A,0}_{a,i}}{rav^0_{[a,i]}p^0} \\
&= \underbrace{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap A_i^1}\frac{n^0_{[a,i,\omega_1]}rav^0_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap A_i^1} n^0_{[a,i,\omega_1]}rav^0_{a,i}}\frac{\phi^{A,0}_{a,i}}{rav^0_{[a,i]}p^0}}_{\mathcal{S}^{A,0}_{A^0,A^1}}\times\underbrace{\frac{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap A_i^1} n^0_{[a,i,\omega_1]}rav^0_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0} n^0_{[a,i,\omega_1]}rav^0_{a,i}}}_{w^{A,0}_{A^0,A^1}} \\
&= \underbrace{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap P_i^1}\frac{n^0_{[a,i,\omega_1]}rav^0_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap P_i^1} n^0_{[a,i,\omega_1]}rav^0_{a,i}}\frac{\phi^{A,0}_{a,i}}{rav^0_{[a,i]}p^0}}_{\mathcal{S}^{A,0}_{A^0,P^1}}\times\underbrace{\frac{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap P_i^1} n^0_{[a,i,\omega_1]}rav^0_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0} n^0_{[a,i,\omega_1]}rav^0_{a,i}}}_{w^{A,0}_{A^0,P^1}} \\
&= \underbrace{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap NT_i^1}\frac{n^0_{[a,i,\omega_1]}rav^0_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap NT_i^1} n^0_{[a,i,\omega_1]}rav^0_{a,i}}\frac{\phi^{A,0}_{a,i}}{rav^0_{[a,i]}p^0}}_{\mathcal{S}^{A,0}_{A^0,NT^1}}\times\underbrace{\frac{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0\cap NT_i^1} n^0_{[a,i,\omega_1]}rav^0_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i^0} n^0_{[a,i,\omega_1]}rav^0_{a,i}}}_{w^{A,0}_{A^0,NT^1}} \\
&= \mathcal{S}^{A,0}_{A^0,P^1}\times w^{A,0}_{A^0,P^1}+\mathcal{S}^{A,0}_{A^0,A^1}\times w^{A,0}_{A^0,A^1}+\mathcal{S}^{A,0}_{A^0,NT^1}\times w^{A,0}_{A^0,NT^1}
\end{aligned}
$$

where

$$
rav^0_{a,i} = (1-\hat{\delta})|a_i^{A,0}-a| + \hat{\delta}\sum_{j\in\mathcal{I}}\pi_j|a_j^{A,0}-a|.
$$

Appendix B

B.1 Theoretical moments details

Here I describe how to compute the variances and covariances needed to calculate the slope between transaction costs and trade size. Let me start with the principal case.

$$cov\Big(10000\frac{\phi^P}{|a^P-a|p}, 100\frac{|a^P-a|}{A}\Big) = \sum_{i\in\mathcal{I}}\sum_{a\in P_i}\frac{n_{[a,i,\omega_1]}}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i}n_{[a,i,\omega_1]}}\Big(10000\frac{\phi^P_{a,i}}{|a^P_i-a|p}-\mathcal{S}^P_{nw}\Big)\Big(100\frac{|a^P_i-a|}{A}-\mathcal{V}^P\Big),$$

$$var(100\frac{|a^P_i-a|}{A}) = \sum_{i\in\mathcal{I}}\sum_{a\in P_i}\frac{n_{[a,i,\omega_1]}}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i}n_{[a,i,\omega_1]}}\Big(100\frac{|a^P_i-a|}{A}-\mathcal{V}^P\Big)^2$$

where $\mathcal{S}^P_{nw}$ is the non-weighted average principal transaction costs and $\mathcal{V}^P$ is the average principal trade size:

$$\mathcal{S}^P_{nw} = \sum_{i\in\mathcal{I}}\sum_{a\in P_i}\frac{n_{[a,i,\omega_1]}}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i}n_{[a,i,\omega_1]}}\Big(10000\frac{\phi^P_{a,i}}{|a^P_i-a|p}\Big)$$

$$\mathcal{V}^P = \sum_{i\in\mathcal{I}}\sum_{a\in P_i}\frac{n_{[a,i,\omega_1]}}{\sum_{i\in\mathcal{I}}\sum_{a\in P_i}n_{[a,i,\omega_1]}}\Big(100\frac{|a^P_i-a|}{A}\Big)$$

For the case of agency trades:

$$cov\Big(10000\frac{\phi^A}{rav\times p}, 100\frac{rav}{A}\Big) = \sum_{i\in\mathcal{I}}\sum_{a\in A_i}\frac{n_{[a,i,\omega_1]}raf_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i}n_{[a,i,\omega_1]}raf_{a,i}}\Big(10000\frac{\phi^A_{a,i}}{rav_{a,i}\times p}-\mathcal{S}^A_{nw}\Big)\Big(100\frac{rav_{a,i}}{raf_{a,i}}\frac{1}{A}-\mathcal{V}^A\Big),$$

$$var(100\frac{rav}{A}) = \sum_{i\in\mathcal{I}}\sum_{a\in A_i}\frac{n_{[a,i,\omega_1]}raf_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i}n_{[a,i,\omega_1]}raf_{a,i}}\Big(100\frac{rav_{a,i}}{raf_{a,i}}\frac{1}{A}-\mathcal{V}^A\Big)^2$$

where $\mathcal{S}^A_{nw}$ is the non-weighted average agency transaction costs, $\mathcal{V}^A$ is the average agency trade size, and $raf_{a,i}$ is the realized agency fraction of customers in state $n_{[a,i,\omega_1]}$ who actually end up trading, i.e., those who hold asset holdings different than their optimal at execution:

$$\mathcal{S}^A_{nw} = \sum_{i\in\mathcal{I}}\sum_{a\in A_i}\frac{n_{[a,i,\omega_1]}raf_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i}n_{[a,i,\omega_1]}raf_{a,i}}\Big(10000\frac{\phi^A_{a,i}}{rav_{a,i}\times p}\Big)$$

$$\mathcal{V}^A = \sum_{i\in\mathcal{I}}\sum_{a\in A_i}\frac{n_{[a,i,\omega_1]}raf_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in A_i}n_{[a,i,\omega_1]}raf_{a,i}}\Big(100\frac{rav_{a,i}}{raf_{a,i}}\frac{1}{A}\Big)$$

$$raf_{a,i} = (1-\hat{\delta})\mathbf{1}_{a^A_i\neq a} + \hat{\delta}\sum_{j\in\mathcal{I}}\pi_j\mathbf{1}_{a^A_j\neq a}.$$

B.2 Regression

Here I present the estimation results for the equation

$$s_{t,d,b} = \alpha + \beta FE + \gamma 100(vol_{t,b,d}/iao_b) + \epsilon_{t,b,d},$$

where $s_{t,b,d}$ is Choi, Huh, and Shin (2023)'s measure of transaction costs Spread1, $vol_{t,b,d}$ is the volume traded, iao_b is the bonds' average amount outstanding, and $FE = [dealer, bond, day]$. The data employed as well as the principal/agency distinction is explained in subsection 5.2.2.

Table B.1: transaction costs - trade size regressions.

Dependent Variable:	Transaction Cost (bp)	
	Principal	Agency
Trade size (pp)	1.45^{***}	0.61^{***}
	(0.13)	(0.12)
Dealer FE	Yes	Yes
Bond FE	Yes	Yes
Day FE	Yes	Yes
Observations	1,505,133	97,305
R^2	0.111	0.019

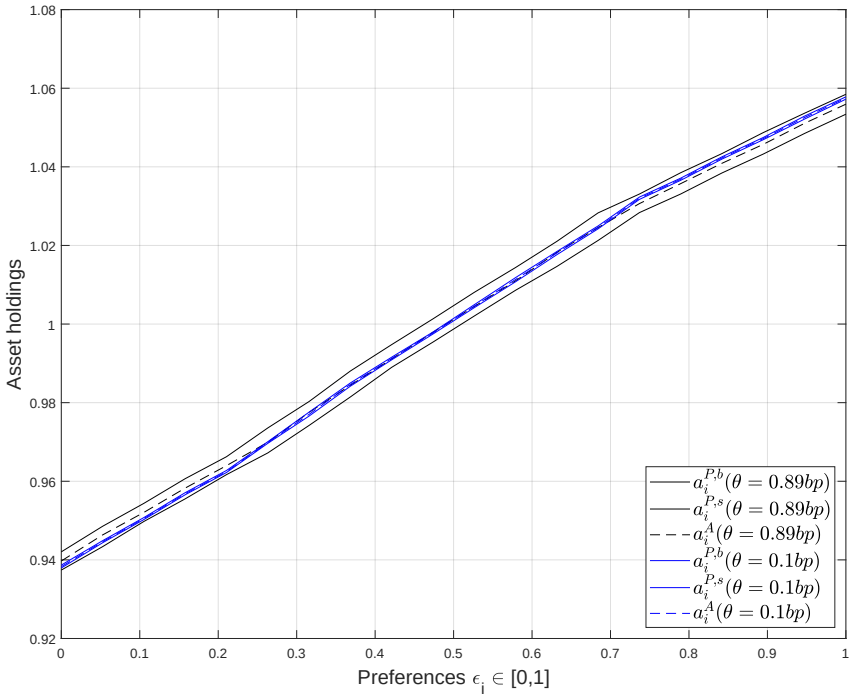
Clustered (Bond & Day) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Appendix C

C.1 Optimal Assets with Low and High Inventory Costs

Figure C.1: Optimal asset as inventory costs increase.



Note: This figure depicts the optimal asset positions of each customer, conditional on her preference type and current holdings, considering $\theta = 0.1$ bps and $\theta = 0.89$ bps. The lower and upper solid lines represent the buyer's and seller's optimal asset holdings under the principal trade, $a^{P,b}$ and $a^{P,s}$, respectively. The dashed line represents the optimal asset holdings under the agency trade, a^{A}. The cases for low and high inventory costs are in blue and black, respectively.

C.2 Quantitative Exercises Robustness Checks

This appendix presents the composition effects (CE) computed for both quantitative exercises, using alternative values of externally calibrated parameters. I consider alternative preference distributions, with $\pi_i \sim Beta(\lambda, \lambda)$, and alternative dealer's bargaining power η. The parameters not affected are kept at their baseline calibration value.

Table C.1: Composition Effects under alternative calibrations

		Composition Effect					
		λ			η		
		0.2	1	5	0.91	0.95	0.99
$\Delta\theta$	CE^P	18.49	32.19	28.65	25.99	32.19	34.58
	CE^A	-0.20	-1.19	0.42	0.50	-1.19	-16.78
$\Delta\beta$	CE^P	79.64	89.54	101.38	74.71	89.54	105.18
	CE^A	-1.14	-1.03	0.26	-1.09	-1.03	-4.08