Composition Effects in OTC Transaction Costs

Mariano J. Palleja

February, 2023.

[Click here for the latest version]

Abstract

The existing literature on over-the-counter markets has documented an increase in the cost of immediate intermediation in the last decade, associated with a shift from principal towards agency trading. This paper argues that the estimates of transaction costs are subject to a composition effect bias. Particularly, the joint determination of trading mechanisms and transaction costs implies that the pool of trades executed in each mechanism before and after a change in market conditions are not comparable. To account for such effect, this paper develops a quantitative search model with heterogeneous customers and trading mechanism choice. Principal trading allows customers to fulfill their trading needs immediately, but is costly. Trading on an agency basis is cheaper, but implies an execution delay. This trade-off defines the optimal mechanism, transaction costs and trading volume for each customer. Whenever market conditions change, some customers migrate from one mechanism to another. If migrating and non-migrating customers pay different costs, the average estimates are biased by composition effects. A calibration to the US corporate bonds market indicates that more than a third of the increase in immediate transaction costs seen in the last decade are due to such bias.

1 Introduction

In over-the-counter (OTC) markets, dealers and customers bargain bilaterally over the costs and execution delays of transactions. To meet these terms, dealers can perform two distinct trading mechanisms: principal or agency. Principal trades are immediate: the dealer uses its inventories to offset the customer's trading needs. Conversely, agency trades imply an expected execution delay, caused by the time it takes the dealer to find a suitable counterparty [1].

This relation between trading mechanisms and execution delays has been largely exploited by the literature to overcome a usual empirical inconvenient: execution delays are not observed. Particularly, when

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

measuring transaction costs, researchers would split trades beforehand according to the trading mechanism used. Principal costs would measure the price of immediacy, whereas agency cost would do so for delayed executions [2].

Although splitting trades in such way purge transaction costs measures from execution delays changes, they overlook the fact that the obtained samples are endogenous: they are the result of a choice. Such endogeneity can be problematic when analyzing the evolution of transaction costs over time, a topic that captured researchers' attention after post-financial crisis regulation and electronification affected OTC markets. Specifically, an estimate of the change in the transaction costs of certain mechanism is subject to a composition bias whenever the samples pre and post change differ. For example, if after a market change the sample of principal trades is composed by customers with higher willingness to pay for immediacy, the impact of such market change on principal costs would be overestimated. This bias can be hardly narrowed whenever the characteristics in which the samples differ cannot be observed.

This paper develops a quantitative search model that explicitly accounts for the composition bias in OTC transaction cost measures. Particularly, I build over the framework in Lagos and Rocheteau (2009) (hereafter LR09). The model features search frictions, heterogeneous risk-averse customers trading a perfectly divisible asset, and Nash bargaining over the terms of trade. My theoretical contribution relative to LR09 is that I allow investors 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 into inventory costs. These features allow me to study the speed-cost trade-off that customers face when choosing a trading mechanism.

I find that, in equilibrium, customer sort across these two mechanisms depending on their liquidity needs. Customers with a larger distance between current and optimal positions chose to trade on principal. Conversely, customers with positions closer to their optimal ones choose to wait for an agency execution. Intuitively, when trading is relatively urgent, the immediacy benefit outweighs the principal premium paid. Given that optimal mechanisms and transaction costs are jointly determined, the different liquidity needs of customers optimally trading in each mechanism are translated into the average transaction costs measured therein. When market conditions change, both trading mechanisms' pools and transaction costs are affected, resembling thus the empirical composition bias in transaction costs this paper targets.

Equipped with the steady-state equilibrium of my model, I study such composition bias. Firstly, I decompose the equilibrium distribution of customers into those that, after a change in the economic

[2] There are two main strategies to identify principal and agency trades. The first one infers agency trades as those offsetting transactions performed the same dealer within a small time window (usually between one and fifteen minutes), labelling as principal all remaining trades (Schultz (2017), Goldstein and Hotchkiss (2020), Choi, Huh and Shin (2021), O'Hara and Zhou (2021)). 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)

environment, remain or not performing trades under the same mechanism, i.e. the non-migrating and migrating customers, respectively. Secondly, for each mechanism I compute measures of transaction cost changes, using both the entire distribution of customers before and after the environment change, as well as the subset of non-migrating customers. The comparison of these measures returns the sign and size of the composition bias.

The model is used to revisit the evidence on transaction costs changes motivated by major changes in the US corporate bond market. In particular, I perform numerical exercises that replicate both the introduction of post 2008 financial crisis 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 composition bias accounts for an economically significant fraction of the changing costs.

In regard to the first exercise proposed, the aftermath of the 2008 financial crisis saw the introduction of new regulations aimed at increasing the financial market resilience. The adoption of the Basel III framework and the Volcker Rule, restrictions meant to reduce banks' exposure to risky assets, negatively affected their dealership activity. Specifically, these new regulations increased the cost of holding assets into bank's 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 over the market transaction costs. The consensus seems to be that principal costs have increased after new regulations took place, with intermediaries shifting away from principal trading towards a larger agency activity (Bessembinder, Jacobsen and Venkataraman (2018), Schultz (2017), Choi, Huh and Shin (2021), Bao, O'Hara, and Zhou (2018), Anderson and Stulz (2017), Dick-Nielsen and Rossi (2019)). I analyze such an increase in inventory costs through the lens of the model. The exercise suggest 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 the traditional voice trading, electronic requests for quotes allow customers to contact multiple dealers at the same time. Dealers benefit from these platforms since their customers' base amplifies, making agency trades a more efficient alternative (Bessembinder, Jacobsen and Venkataraman (2018), O'Hara and Zhou (2021)). Collectively, the evidence points out that the rise of electronic trading had a similar effect as the new regulations, promoting a shift from principal towards agency trading. I revisit this evidence by reducing the execution delays of the model. I find that the composition bias explains most of the change in principal costs, while it implies a negligible underestimation of the change in agency costs.

1.1 Related Literature

This paper develops a theoretical model of trading mechanism choice to evaluate quantitatively the composition bias present in OTC transaction costs measures. It contributes to three strands of the literature.

In first instance, this paper contributes to the search literature in OTC markets, pioneered by Duffie, Gârleanu and Pedersen (2005) and Lagos and Rocheteau (2009). 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.

This paper also contributes to the theoretical literature that explicitly accounts for principal and agency trading in OTC markets (Cimon and Garriot (2019), Saar et. al. (2020), An (2020) and An and Zheng (2020)). My contribution to this research agenda is twofold. On the one hand, this paper is the first to address OTC trading mechanisms as a customers' choice [3]. While the existing literature focuses on the dealer's mechanism decision as a function of their inventory costs, here I model customers' mechanism choice as a function of both the (translated) inventory costs faced by dealers and their own liquidity needs. Therefore, this paper is able to speak about the endogenous response of customers whenever the market environment changes. On the other hand, unlike the existing literature, my model features two characteristics necessary to study composition effects in a quantitative way. Firstly, both the optimal trading mechanisms and the transaction costs are functions of each specific customer liquidity needs. Secondly, these liquidity needs result from the combination of each customer's preferences and the distance between their current and optimal positions. Jointly, these two ingredients provide a non-degenerated distribution of transaction costs within each trading mechanism, which I use to measure transaction costs as empirical researchers would.

Finally, this paper complements the empirical literature that address transaction costs changes and trading mechanisms 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. Particularly, researchers have shown that principal trading is less abundant and more costly (Bessembinder, Jacobsen and Venkataraman, 2018; Schultz, 2017; Choi, Huh and Shin, 2021; Bao, O'Hara, and Zhou, 2018; Anderson and Stulz, 2017; Dick-Nielsen and Rossi, 2019). Additionally, the empirical evidence indicates that the rising electronic venues had attracted volume towards agency trading, reducing the cost of such trades (Bessembinder, Jacobsen 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 et.al., 2021). A common feature across these papers is the lack of customers' data, which prevents

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

them from controlling the documented customers' endogenous migration when computing transaction costs changes [4]. I complement these papers by informing about the size and sign of the consequent composition bias. To achieve this goal, I exploit the model to construct counterfactual distributions in which transaction costs can be measured using a steady sample of customers. I show that the transaction costs' estimates 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 describe the features of the model. I start by narrating the general environment, and the problems that both customers and dealers face. Later I show how terms or trade are settled, 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 are two types of infinitely lived agents: customers and dealers, both in unite 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*. The asset is perfectly divisible and in fixed supply $A \in \mathbb{R}_+$. Fruit is perishable, non-tradable, and it is produced by the asset in 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. The utility function $u_i(a)$ is twice continuously differentiable, strictly increasing, strictly concave, and satisfies Inada conditions. 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 is the trade motive in the model, and can be interpreted as changing hedging needs, investing opportunities, etc.

Customers can trade assets only when they are contacted by a dealer, an event that is governed by a Poisson process with arrival rate of α. Once a customer meets a dealer, she chooses among two kind 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 his excess position at the inter-dealer price p and pays an intermediation fee of ϕ^P. On the other hand, if she opt for an agency trade, she waits until the dealer finds her a counterparty, and meanwhile enjoys the utility provided by her current asset

[4] 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 customers characteristics, they use bond and trade characteristics to predict the optimal mechanism of a trade.

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 re-balances her position at p and pays the dealer a fee 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 be contacted by a dealer or waiting for their agency trade to be executed. These two states will be denoted by ω_1 and ω_2, respectively.

Fees and quantities in both kinds of trades are determined through a Nash bargaining protocol, that takes place at the moment of contact with the dealer. This timing assumption implies that, in the agency protocol, 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 cycle from the moment she contacts a dealer until she execute her trade.

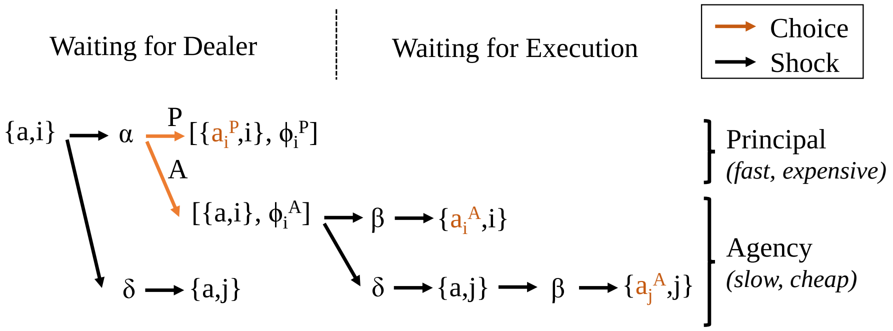

Figure 1. Customer Path. This figure shows a customer's path through the state space. Shocks are depicted by black arrows, and include the contact with dealers (α), the change of preference (δ), and the execution of the agency trade (β). The customer's choice are depicted in orange arrows and include optimal trading mechanism and the corresponding new asset holdings.

The maximum expected discounted utility attainable by a customer waiting for a dealer with type i and asset holding a at time t, $V_i(a, t)$, satisfies

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

where

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

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 $k(s)$, conditional on the customer being type $k(t) = i$ at t. Fees and prices are expressed in units of the *numéraire* good.

Note that the optimal asset holdings under the two trading protocols, $a^P_{k(T_\alpha)}$ and $a^A_{k(T_\beta)}$, might differ for two reasons. Firstly, a customer might change her type during the waiting period of a delayed trade. Hence, types $k(T_\alpha)$ and $k(T_\beta)$ might be different. Secondly, the fees 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.

In turn, a representative dealer does not hold positions and her instantaneous utility equals her consumption of the *numéraire* good d. Thus, her expected utility is given by the present value of the fees she collects net of the costs she incurs. Dealers trade on behalf of their customers in the inter-dealer market. If they are asked to execute the trade immediately, they need to face a cost, which is denoted by the function f. It is assumed that f is positive, strictly increasing and continuous in the trade size. This reduced form specification is designed to represent both the inventory cost a dealer incurs when performing principal trades to meet immediacy demand and the cost of searching and trading the asset (potentially at a premium/discount) in case she does not hold it[5]. On the other hand, if the client asks the dealer to perform an agency trade, they wait until finding a counterparty and the fee is charged when the transaction

[5] In terms of modeling choice, this reduce form formulation allows to draw a link between demand for immediacy and dealers' balance sheet pressure without dealing with inventories as an additional state variable. See Cohen et.al. (2021) for a search model with explicit inventory in OTC markets.

is executed. A dealers' maximum expected discounted utility satisfies

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

$$\text{where} \qquad \Phi_i(a, T_\gamma) = \begin{cases} \phi_i^P(a, T_\gamma) - f\left(a_i^P - a\right) & \text{if P trade} \\ \phi_i^A(a, T_\gamma) & \text{if A trade} \end{cases}$$

and $T_\gamma = min\{T_\alpha, T_\beta\}$ and the integration over the probability measure H_{T_γ} 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 allocations, their corresponding intermediation fees, and the optimal trading mechanisms. I find that, in equilibrium, customers sort across mechanisms depending on their liquidity needs.

2.2.1 Asset Allocations and Intermediation Fees

Once a customer contacts a dealer and chooses a trading mechanism, optimal asset holdings and fees are chosen as the outcome of a Nash bargaining problem, where the dealer has η bargaining power. When trading at 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 intermediation fee minus the cost of performing principal trades. Hence, the Nash product writes

$$\{a_i^P(a,t), \phi_i^P(a,t)\} = \arg\max_{(a',\phi')} \Big\{V_i(a',t) - V_i(a,t) - p_t[a'-a] - \phi'\Big\}^{1-\eta}\Big\{\phi' - f(a'-a)\Big\}^{\eta}.$$

The solution for optimal principal terms of trade is

$$\phi_i^P(a,t) = \eta\big[V_i(a_i^P(a,t),t) - V_i(a,t) - p_t[a_i^P(a,t) - a]\big] + [1-\eta]\big[f(a_i^P(a,t) - a)\big], \tag{3}$$

$$a_i^P(a,t) = \arg\max_{a'} \quad V_i(a',t) - V_i(a,t) - p_t[a'-a] - f(a'-a). \tag{4}$$

Let me impose more structure on the cost function faced by dealers when trading immediately. Particularly, I assume a symmetric linear specification, $f(a_i^P - a) = \theta \times p_t |a_i^P - a|$, where $\theta \in [0, \frac{r}{r+\beta})$ is the constant marginal cost per (*numeraire*) dollar traded [6]. This assumption will have two important consequences for principal trades.

[6] Modeling an appropriate functional form for inventory costs is not a trivial task. Such function should probably include deviations from an optimal inventory target, feature out of the scope of this paper.

Firstly, the principal problem is greatly simplified. Conditional on the trade direction, i.e buying or selling on principal, the cost function becomes linear and so customers choose their optimal holdings independently of their current positions. In this regard, trading on principal can be thought as a two step procedure. Initially the customer sells his current holdings a, loosing $V_i(a,t) - ap(t)$, and later she buys her optimal holdings a', obtaining $V_i(a',t) - a'p(t) - f(a'-a)$. Without the inclusion of immediacy costs, current asset holdings would not be present in the second step, and so they could be ignored when solving the optimization problem. A piece-wise linear immediacy cost function yields the same result, making the optimal principal asset holdings, conditional on trade direction, independent of the current position. Consequently, inventory costs can be translated into an increase in the effective price a customer pays when buying or into a decrease in the effective price a customer obtains when selling.

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[7]. Whenever the gain in lifetime utility minus the inter-dealer price paid for such trade does not outweigh the immediacy costs, is better not to trade on a principal basis. Furthermore, if keeping the current position is preferred over engaging into an agency trade, the optimal policy is not to trade at all.

These two consequences can be easily seen replacing the cost function into eq. (4), and optimizing conditional on trade direction. Particularly, equation (4) can be split into three regions, within which optimal asset holdings can be easily characterized [8]:

- No trade region, $NoT_i : a \,|\, [V_i(a',t) - a'p(t)] - [V_i(a,t) - ap(t)] \le \theta p_t \times |a'-a| \quad \forall a' \neq a,$
  $a_i^P(a,t) = a.$

- Buy region, $Buy_i : a \,|\, [V_i(a',t) - a'p(t)] - [V_i(a,t) - ap(t)] > \theta p_t \times |a'-a| \quad \text{for some } a' \in (a,\infty)\,,$
  $a_i^P(a,t) = a_i^{P,b}(t) = \arg\max_{a'}\{V_i(a',t) - p(t)[1+\theta]a'\}.$

- Sell region, $Sell_i : a \,|\, [V_i(a',t) - a'p(t)] - [V_i(a,t) - ap(t)] > \theta p_t \times |a'-a| \quad \text{for some } a' \in [0,a)\,,$
  $a_i^P(a,t) = a_i^{P,s}(t) = \arg\max_{a'}\{V_i(a',t) - p(t)[1-\theta]a'\}.$

[7] Given that most of the databases are based on transaction data, the empirical evidence related to no trades is hard to find. Hendershott et. al. (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, prevents us from reading these numbers through the lens of the present model.

[8] These three sets of asset holdings are guarantee to be convex if the value function is strictly concave. In such case, optimal holdings can be represented as:

$$a_i^P(a,t) = \begin{cases} a_i^{P,b} & \text{if } a < a_i^{P,b} \\ a & \text{if } a_i^{P,b} \le a \le a_i^{P,s} \\ a_i^{P,s} & \text{if } a > a_i^{P,s}. \end{cases}$$

However, the trade mechanism decision implies that the value function has kinks, and so I cannot proof the convexity of the aforementioned sets. I check numerically both the convexity of the sets as well as the pattern of the optimal holdings and they both hold robustly.

In turn, agency trades implies an expected execution delay, during which the customer might suffer a preference shock. Hence, a specific timing assumption regarding when optimal holdings and fees are set is needed. To allow for a tractable state space, it is assumed that fees are arranged when customers and dealers meet, but that optimal holdings are decided at execution. In other words, agency fees will be settled based on the expected gains from trade of customers who re-balance their positions according to their preference type at execution[9]. 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 fee collected. Terms of trade when agency is chosen are set according to

$$\begin{aligned}
&\{[a^A_{i(T_\beta)}(T_\beta)]^I_{i=1}, \phi^A_{i(t)}(a,t)\} \\
&= \arg\max_{\{a''_i\}^I_{i=1}, \phi''} \Big\{\mathbb{E}_{i,t}\Big[\int_t^{T_\beta} e^{-r[s-t]} u_{k(s)}(a) ds + e^{-r[T_\beta - t]}\big[V_{k(T_\beta)}(a''_{k(T_\beta)}, T_\beta) - p_{T_\beta}[a''_{k(T_\beta)} - a] - \phi''\big]\Big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad - V_i(a,t)\Big\}^{1-\eta}\Big\{\mathbb{E}_t\big[e^{-r[T_\beta - t]}\phi''\big]\Big\}^{\eta}.
\end{aligned}$$

The optimal terms in the agency trade are

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

$$a^A_{i(t)}(t) = \arg\max_{a''} \quad \{V_{i(t)}(a'', t) - p_t[a'' - a]\}. \qquad (6)$$

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 bargain power but smaller contact rate $\kappa = \alpha(1 - \eta)$. In other words, for a customer it is equivalent to contact dealers with bargain power η at contact rate α, than to contact dealers with null bargain power but less often, at contact rate κ. I therefore refer to κ as the bargain-adjusted contact rate. Second, I use analytical expressions for all the expectation related to the shocks of the model[10]. Finally, since I am going to focus on the steady state solution of the model, I ease

[9] An alternative modeling choice is to assume that customers and dealers commit upon contact to trade 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 will require to track 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 optimal allocations and fees 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 nor in the consequent terms of trade.

[10] See the Appendix for a step by step computation.

exposition by replacing $p_t = p$.

$$\begin{aligned} V_i(a) = \bar{U}_i^{\kappa}(a) + \hat{\kappa}\big[[1-\hat{\delta}^{\kappa}] \max\big\{V_i(a_i^P) - p[a_i^P - a] - f(a_i^P - a), \bar{U}_i^{\beta}(a) + \hat{\beta}[\bar{V}_i^A - p[\bar{a}_i^A - a]]\big\} \\ + \hat{\delta}^{\kappa} \sum_j \pi_j \max\big\{V_j(a_j^P) - p[a_j^P - a] - f(a_j^P - a), \bar{U}_j^{\beta}(a) + \hat{\beta}[\bar{V}_j^A - p[\bar{a}_j^A - a]]\big\}\big], \qquad (7) \end{aligned}$$

where

$$\begin{aligned} &\bar{U}_i^{\nu}(a) = \Big[[1-\delta^{\nu}]u_i(a) + \delta^{\nu}\sum_j \pi_j u_j(a)\Big]\frac{1}{r+\nu} \\ &\bar{V}_i^A = [1-\delta^{\beta}]V_i(a_i^A) + \delta^{\beta}\sum_j \pi_j V_j(a_j^A) \quad , \quad \bar{a}_i^A = [1-\delta^{\beta}]a_i^A + \delta^{\beta}\sum_j \pi_j a_j^A \\ &\hat{\kappa} = \frac{\kappa}{r+\kappa} \quad , \quad \hat{\beta} = \frac{\beta}{r+\beta} \quad , \quad \hat{\delta^{\nu}} = \frac{\delta}{r+\delta+\nu} \quad , \quad \nu = \{\kappa, \beta\}. \end{aligned}$$

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 his preferences, and so this term is a convex combination of the utility under type i and under 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. In the first case, the trading mechanism choice varies according to the different utility gains that the customer of type i expects to get with the principal and with the agency protocols. In the second case, the choice is a function of the expected utility the expected type would get under the two trades. Thus, given the assumption of no serial correlation among preference shocks, this later case is independent of current type i, and it is solely a function of current asset holdings a.

If customers choose to trade on principal, the execution is immediate. 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.

Note that the agency trade presents a similar structure as the scenario of a customer who is waiting to contact a dealer. However, there are two main difference between trading protocols. 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 immediacy costs that dealers perform using intermediation fees. These two differences define the trade-off that customers will have to solve to choose which kind of trade to perform.

2.2.2 Trading Mechanism Choice

When a dealer is contacted, the customer must choose between an immediate principal trade or a delayed agency trade. As usual when solving for binary choice problems, it is convenient to analyze indifference conditions. Given that customers who contacted a dealer are characterized by two state variables, preference type and asset holdings, I will look for the current asset holding thresholds that, conditional on type, make each customer indifferent among trading protocols, $\hat{a}_i$. The indifference condition for a type i customer who contacts a dealer at time t is given by:

$$V_i(a_i^P) - p[a_i^P - \hat{a}_i] - f(a_i^P - \hat{a}_i) = \bar{U}_i^{\beta}(\hat{a}_i) + \hat{\beta}\big[\bar{V}_i^A - p[\bar{a}_i^A - \hat{a}_i]\big]. \tag{8}$$

Firstly, consider Eq.(8) for the cases where optimal asset holdings under principal trading are independent of current asset holdings. This happens when customers would trade if they were to opt for the principal protocol, i.e. the Buy_i and $Sell_i$ regions. The indifference condition can be expressed as:

$$V_i(a_i^P) - p[a_i^P - \hat{a}_i] - \psi\theta p[a_i^P - \hat{a}_i] = \bar{U}_i^{\beta}(\hat{a}_i) + \hat{\beta}\big[\bar{V}_i^A - p[\bar{a}_i^A - \hat{a}_i]\big], \tag{9}$$

where $\psi = 1$ if the customer is buying, $a \in Buy_i$, and $\psi = -1$ if she is selling, $a \in Sell_i$. As previously stated, conditional on increasing or reducing positions, principal trade gains become linear in current holdings a. As a consequence, the gains from a principal trade increase at a constant rate in a. This can be seen in the left hand side of the indifference equation. On the other hand, in the agency protocol, the customer keeps his current asset holdings until some counterparty is found. Firstly, she derives utility according to $\bar{U}_i^{\beta}(a)$, which is a convex combination of marginally decreasing functions. 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 eq. (6)). Therefore, the gains from a delayed inter-mediated trade are marginally decreasing in a. I will exploit these differences in the two type of trades to find the current asset holdings thresholds as the roots of equation (9). Let us rearrange the arguments of such indifference equation:

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

The left hand side, B_i, is independent of current asset holdings a, while the two arguments in the right hand side are not. Firstly, $C_i(\hat{a}_i)$ is a twice continuously differentiable, strictly increasing, and strictly concave function that satisfies Inada conditions in current asset holdings a. Secondly, $D(\hat{a}_i)$ 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. The sign of $D(\hat{a}_i)$ will define the shape of the right hand side. Given the assumption made about the constant marginal cost parameter, $\theta < \frac{r}{r+\beta}$, $D(\hat{a}_i)$ is a decreasing linear function on a, and the right hand side is thus inverse U-shaped [11].

Define $\hat{a}_i^{h,\rho}$, with $h = \{1, 2\}$ and $\rho = \{b, s, nt\}$, as the current asset holdings that make customer 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 Γ_i^P and Γ_i^A, with $\Gamma = \{Buy, Sell, NoT\}$, as the type specific sets 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 trade direction. Eq. (9) provides two possible scenarios for each trading direction:

Principal Buyers:

$\bullet B_i \geq C_i(\hat{a}_i) + D_i(\hat{a}_i) \quad \forall \hat{a}_i :$ $\qquad Buy_i^P = Buy_i.$

$\bullet B_i < C_i(\hat{a}_i) + D_i(\hat{a}_i)$ for some $\hat{a}_i :$ $\quad Buy_i^P = Buy_i \cap \{[-\infty, \hat{a}_i^{1,b}] \cup [\hat{a}_i^{2,b}, \infty)\}$, $Buy_i^A = Buy_i \setminus Buy_i^P.$

Principal Sellers:

$\bullet B_i \geq C_i(\hat{a}_i) + D_i(\hat{a}_i) \quad \forall \hat{a}_i :$ $\qquad Sell_i^P = Sell_i.$

$\bullet B_i < C_i(\hat{a}_i) + D_i(\hat{a}_i)$ for some $\hat{a}_i :$ $\quad Sell_i^P = Sell_i \cap \{[-\infty, \hat{a}_i^{1,s}] \cup [\hat{a}_i^{2,s}, \infty)\}$, $Sell_i^A = Sell_i \setminus Sell_i^P.$

Let us consider now Eq.(8) for the cases where customers would not trade if they were to opt for the principal protocol.

$$V_i(\hat{a}_i) = \bar{U}_i^{\beta}(\hat{a}_i) + \hat{\beta}\big[\bar{V}_i^A - p[\bar{a}_i^A - \hat{a}_i]\big]. \tag{10}$$

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.

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

The left hand side, B_i, is still independent of current asset holdings a_i. Regarding the right hand side, $D(\hat{a}_i)$ is linear and strictly increasing in a. In turn, $C_i(\hat{a}_i)$ subtracts from a strictly increasing and strictly concave function a function $V_i(\hat{a}_i)$ that, at this point, is unknown. The shape of $C_i(\hat{a}_i)$ determines the region

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

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 $C_i(\hat{a}_i) + D_i(\hat{a}_i)$ is U-shaped, and so the region under which a customer with specific preference i decides not to trade follows:

Principal Non Traders:

$\bullet B_i < C_i(\hat{a}_i) + D_i(\hat{a}_i) \quad \forall \hat{a}_i : \qquad NoT_i^P = \emptyset.$

$\bullet B_i \geq C_i(\hat{a}_i) + D_i(\hat{a}_i) \quad \text{for some } \hat{a}_i : \quad NoT_i^P = NoT_i \cap \{[\hat{a}_i^{1,nt}, \hat{a}_i^{2,nt}]\,,\, NoT_i^A = NoT_i \setminus NoT_i^P.$

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 $\{a_t, i_t, \omega_t\}$. Thus I firstly develop the equations needed to compute the steady state distribution $H(a, i, \omega)$ over such individual states. Secondly, 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 $a_t \in R_+$ might be included into the ergodic set. In order to have a tractable discrete state space, the calibrations are going to be restricted to those where $\cap_{i=1}^{I} NoT_i^P = \emptyset$. In other words, I impose that there is no asset position such that every type decides not to trade when holding it. Under this restriction, given that $\pi_i > 0\, \forall i$, every customer with any asset holdings will eventually trade. Hence, in equilibrium, 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 $\{\Gamma_i^P, \Gamma_i^A\}_{i=1}^{I}$, $\Gamma = \{Buy, Sell, NoT\}$, indicate which kind of trade customers wish to perform. These policy function 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 compute the stationary distribution of this model.

$$n_{[a_i^{P,b},i,\omega_1]}: \quad \delta\pi_i \sum_{j\neq i} n_{[a_i^{P,b},j,\omega_1]} + \alpha \sum_{a\in Buy_i^P} n_{[a,i,\omega_1]} = n_{[a_i^{P,b},i,\omega_1]} \big[\delta[1-\pi_i] + \alpha \mathbf{1}_{[a_i^{P,b}\notin NoT_i^P]}\big] \tag{11}$$

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

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

$$n_{[a_j,i,\omega_1]}: \quad \delta\pi_i \sum_{j\neq i} n_{[a_j,j,\omega_1]} = n_{[a_j,i,\omega_1]} \big[\delta[1-\pi_i] + \alpha \mathbf{1}_{[a_j\notin NoT_i^P]}\big], \quad a_j \in \cup_{j\neq i}\{a_j^{P,b}, a_j^{P,s}, a_j^{A}\} \tag{14}$$

$$n_{[a_i,i,\omega_2]}: \quad \delta\pi_i \sum_{j\neq i} n_{[a_i,j,\omega_2]} + \alpha n_{[a_i,i,\omega_1]} \mathbf{1}_{[a_i\in\Gamma_i^A]} = n_{[a_i,i,\omega_2]} \big[\delta[1-\pi_i] + \beta\big], \quad a_i \in \mathcal{A}^* \tag{15}$$

The left hand side of these equations represent the inflow in an specific individual state, and the right hand side represents the outflow. As Figure 1 shows, in any given time lapse in which a unique shock has hit the economy, three kinds of forces might have moved customers across states. Let us first consider the preference shock. The mass of customers of an individual state with preference i increase whenever customers from other states, with the same asset holdings and in the same waiting stage, receive the preference shock i. This happens with instantaneous probability $\delta\pi_i$. Similarly, that mass of customers decreases whenever customers therein are hit by preference shocks others than i. This happens with instantaneous probability $\delta(1-\pi_i)$. Secondly, lets consider the contact with dealer shock. This shock is only received by people waiting for a dealer, i.e by customers within states where $\omega = \omega_1$, and happens with instantaneous probability 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 opt 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 it will remain in his current state until a preference shock eventually hits her. Finally the execution shock, which happens with instantaneous probability β, needs to be addressed. Obviously, such shock is only received 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 optimal holdings at the moment of execution, this shock will move customers towards the state in which optimal agency asset holdings correspond with their preference type.

The set of equations (11)-(15) 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 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 and Lucas (1987) provides the conditions for such convergence result [12].

Once solved for the stationary distribution, the market clearing equation can be computed, and thus the steady state equilibrium price p can be founded. 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:

$$\sum_{h=1}^{2}\sum_{i=1}^{I}\sum_{a\in\mathcal{A}^*} a n_{[a,i,\omega_h]} = A. \tag{16}$$

Equation (16) highlights the fact that, in spite of the two different trading mechanisms, the market is not segmented. All trades are cleared in the inter-dealer market. Therefore, in every instant, the excess of supply in one mechanism is compensated by the excess of demand in the other.

2.4 Equilibrium

An equilibrium for this model is defined as a list of optimal asset holdings $\{a_i^P(a), a_i^A\}_{i=1}^I$, fees $\{\phi_i^P(a), \phi_i^A(a)\}_{i=1}^I$, 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 (11)-(15), and p satisfies (16).

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 [13]. To solve for the partial equilibrium steady state of the model, I rely on the value function iteration method, enhanced with Howard's improvement step [14]. This procedure returns the policy and value functions conditional on any given price p. In turn, these functions are nested within the computation of the market clearing condition, which solves for the general equilibrium steady state. 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 $V_i(a)$

[12] Basically, there should exist at least one state which receives inflows from all states with strictly positive probability, whether directly or indirectly (through the iteration of T). A sufficient condition for this to happen is that it exist a type i and a type j such that $[a_i^{P,b}, a_i^{P,s}, a_i^A] \in [\Gamma_j^P \cap \Gamma_j^A] \setminus NoT_j^P$, i.e. such that all optimal holdings of some i leads to choose the same trading mechanism and trading direction when turn into type j. Firstly, $\pi_i > 0\,\forall i$ and $\delta \in (0,1)$, therefore all types can be reached through preference shocks. Secondly, after any trade execution, customers go back to the waiting for a dealer stage, $\omega = \omega_1$. Finally, the existence of a type i for which all his optimal policy functions would make him follow the same trading protocol under type j guarantees that such individual state would receive inflows directly or indirectly from all individual states.

[13] LR09, Online Appendix D, show how the value function can be decomposed into a sequence of tractable expected gains from trade. This allow the authors to get a closed form solutions.

[14] See Appendix A.4 for the necessary and sufficient conditions to use value function iteration as solution method.

(b) Compute optimal asset holdings $\{a_i^P(a), a_i^A\}_{i=1}^I$ using Eq.(4) and Eq.(6).

(c) Compute trading mechanism choice for each pair $\{i, a\}$, using Eq.(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 Eq.(7).

ii. Compute trading mechanism choice for each pair $\{i, a\}$, using Eq.(8)

(e) Update $V_i(a)$ using Eq.(7) until convergence.

2. Define trading mechanism sets $\{\Gamma_i^P, \Gamma_i^A\}_{i=1}^I$ using Eq.(9) and Eq.(10).

3. Compute transition matrix T using Equations (11), (12), (13), (14) and (15).

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 eq. (16) converges towards zero.

3 Equilibrium Allocations

In the following subsections I study the policy functions of the model. For this, I initially outline the calibration used, and later map customers preferences and holdings with their optimal allocations and corresponding fees payed.

3.1 Calibration

The calibration used closely follows the baseline parameters in LR09, and aims to replicate stylized facts about the US secondary market for corporate bonds. I normalize the supply of assets at $A = 1$, set the time length at a day and the discount rate at 7% per year. In line with LR09, I assume customers contact dealers, on expectation, on a daily basis, and fix $\alpha = 1$. I follow Hugonnier, Lester and Weill (2020) and set the bargaining power of dealers at $\eta = 0.9$, value that implies an expected bargain-adjusted contact every 10 days. I assume customers suffer preference shocks at the same rate at which they can contact dealers, and fix $\delta = 1$, which yields a daily turnover of 1.07% [15]. Preferences are modeled with CRRA utility functions. Particularly, $u_i(a) = \epsilon_i \times a^{1-\sigma}/(1-\sigma)$, where the risk aversion parameter is fixed at $\sigma = 2$ and the support for the values of ϵ_i is $\left\{\frac{i-1}{I-1}\right\}_{i=1}^I$ with $I = 20$. The probabilities assigned to each preference type are assumed to be uniformly distributed, $\pi_i = 1/I \; \forall i$.

[15] This measure is compatible with those bonds that are more actively traded. Using data from TRACE for years 2016-2017, the average monthly turnover is computed at 4%, with a standard deviation of 5 percentage points.

Compared to LR09, two new parameters are introduced: the marginal cost per dollar traded on a principal basis, θ, and the (inverse) of the expected execution delay of an agency trade, β. These two parameters lack of related observable measures, therefore I need to develop a strategy for their calibration.

Regarding θ, I focus on the regulation-induced costs dealers face when including assets into their inventory. This approach allows me to draw a direct link between the changing regulation and the composition effect present in transaction costs estimates. According to Greenwood et. al. (2017) and Duffie (2018), the leverage ratio requirement (LRR) is the most tightly binding constraint for most U.S. banks after the new regulations were set. This regulation requires banks to hold capital for an amount of 3-5% of the non-risk-weighted value of assets in inventory[16]. Restricting the attention to this most binding regulation, the inventory cost faced by a dealer buying $p(a' - a)$ worth of assets, offloading this position after m days, facing $x\%$ of capital requirement and incurring into an opportunity cost of $z\%$, is $p[a' - a][e^{zm} - 1]x\%$. The model counterpart of such round-trip principal trade cost is $2\theta p[a' - a]$. Therefore, the following mapping is obtained: $\theta = [e^{zm} - 1]x\%/2$. The average holding period m is taken from Goldstein and Hotchkiss (2020), and equals 10.6 days. I further assume that the opportunity cost of capital held z equals the discount rate r. In order to recreate a scenario were post 2008 financial crisis regulations were not yet settle, I fix the baseline capital requirement percentage in x=1%. Therefore, the marginal cost face by dealers when trading on principal is $\theta \approx 0.1 b.p.$.

The second parameter introduced, β, accounts for the expected execution delay of an agency trade. The available data only inform us of when trades are executed, but not on the initial customer-dealer contact that started the transaction. Hence, β needs to be calibrated based on indirect observable measures. Among many objects of the model that are affected by this parameter, I choose to target the ratio between average transaction costs of principal and agency trading. To a first order, an increase in β implies that agency execution is faster, and so both agency trading surplus and fees payed increase. This reduces the ratio of principal over agency costs. To a second order, the execution rate affects the optimal trading mechanism choice, and so the steady state distribution over which average transaction costs are computed. As it will be explained in subsection 5.2, the overall effect of an increase in β is to decrease the ratio between average transaction costs of principal and agency trading. To obtain such ratio, I compute intermediation costs series for IG bonds from 2006 to 2018. Consistent with the model, these series are calculated as the percentage deviation between customer-dealer prices and same bond-day inter-dealer prices. I find an average ratio of roughly 2, which is approximated by $\beta = 1/3$ [17]. The choice of targeting the transaction cost ratio follows from the goal of this paper, which is to asses the composition effect present in transaction costs measures. As was previously depicted, the composition effect depends on the differential costs payed by migrating and non

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

[17] The execution delays are in line with Wu (2022), who recovers latent trading delays for the US corporate bond market from a structural model that links such figures with observable liquidity premiums.

migrating trades. The ratio between average transaction costs of principal and agency plays a determinant role in such cost differential, and thus needs to be properly matched.

Table 1: Calibrated Parameters. Unit of time = 1 day

Parameter	Description	Value
A	Asset supply	1
r	Time discount	$ln(1.07)[360\frac{5}{7}]^{-1}$
σ	CRRA coeff	2
$1/\alpha$	Expected days to contact dealer	1
$1/\delta$	Expected days to receive pref. shock	1
$1/\beta$	Expected days to execute agency trades	3
η	Dealer's bargain power	0.9
θ	Principal trades cost	$0.1b.p.$

3.2 Who Trades In Each Mechanism and at Which Cost?

In this section I present the equilibrium allocations under the proposed calibration. The goal is to characterize the pool of trades in each mechanism, and to show how these characteristics are translated into the intermediation costs customers pay.

The policy functions are presented in Figure 2. For each type $i \in [1, I]$ I compute both the optimal asset holdings conditional on trading mechanism and the trading mechanism choice. Regarding the optimal holdings, the lower and upper solid lines represent the buyer's and seller's optimal allocations 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, if holding $a > a^{P,s}$ would be a seller, and with current assets $a \in [a^{P,b}, a^{P,s}]$ would not trade on principal. The distinction of these three regions is a direct consequence of the inclusion of inventory costs. On the one hand, in the principal market, 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 trade surplus of those customers with current holdings between the buyer's and seller's optimal allocations 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 fee 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 that of the principal traders.

Figure 2 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, under the proposed calibration, every customer for which the principal trading cost is higher than their trade surplus (customer located between the two solid lines) 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.

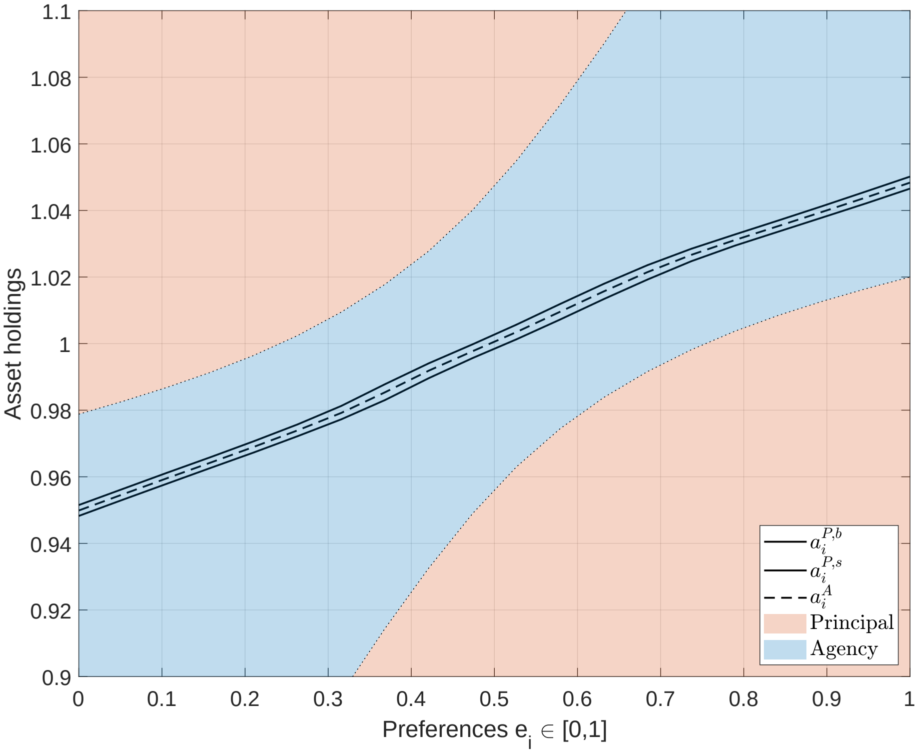

Figure 2. Optimal asst holdings and trade choice. This figure depicts the policy functions of each customer, conditional on her preference type and current holdings. The lower and upper solid lines represent the buyer's and seller's optimal allocations under the principal trade, $a^{P,b}$ and $a^{P,s}$, respectively. The dashed line represents the optimal allocation under agency trading, a^A. Regarding the trading mechanism choice, the principal and agency regions are shaded in orange and blue, respectively.

To better understand these policy functions, consider for example customers with preferences $e_i = 0.6$. When contacted by a dealer, these customers compute what their optimal allocations would be as principal traders, $a_i^{P,b}$ or $a_i^{P,s}$, and what they would expect to trade after the waiting period of the agency trade, $\bar{a}_i^A$. Given these optimal allocations, they evaluate, using eq (8), which trade to perform. As Figure 2 shows, customers owning roughly less than 0.98 units of the asset perform a principal buy. Customers holding between 0.98 and 1.08 units perform an agency trade. Finally, customers with type $e_i = 0.6$ holding assets above 1.08 choose to sell on a principal basis.

Figure 2 reveals a key feature regarding trading pools: customers choosing agency trades are concentrated in the center of the preference-assets state space. Conditional on preference types, agency trading is mostly performed by customer with current asset holdings close to their optimal allocations. Moreover, conditional on the asset holdings, agency is mostly performed by customers with preferences close to the mean. These two features can be understood analyzing the indifference condition (8). This equation highlights the speed-cost trade-off that customers face when choosing a trading mechanism. Particularly, customers compare the utility lost implied by a delayed execution with the utility lost implied by the payment of the principal cost. Consider first customers holding the same preference type. The larger the distance between current and optimal holdings is, the bigger the marginal trade surplus per unit exchanged. This is a consequence of modeling risk averse agents. On the other hand, the cost dealers face per unit traded is constant. Hence, whenever current asset holdings are far away from the optimal allocations, the average cost/surplus ratio of trading on principal decreases and is outweighed by the utility lost implied by delaying the reallocation. A customer with an extreme asset position would then be more likely to trade on principal. Conversely, a customer with a position close to her optimal allocation would be prone to trade on an agency basis. Secondly, consider customers holding the same asset position. If such position is relatively small (big), given that optimal allocations are increasing in preference types, customers with high (low) preference types will be far away from their optimal holdings. Hence, customers with extreme preferences will find themselves more often far away from their optimal holdings than customers with moderate preferences. Given the relation about trading mechanism choice and distance between current and optimal position, the model tell 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.

I next present the distribution of intermediation fees 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, an intermediation fee is a convex combination between the customer's expected trading surplus and the dealer's trading 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 3, which maps intermediation fees with the asset-preference state-space, depicts such heterogeneity.

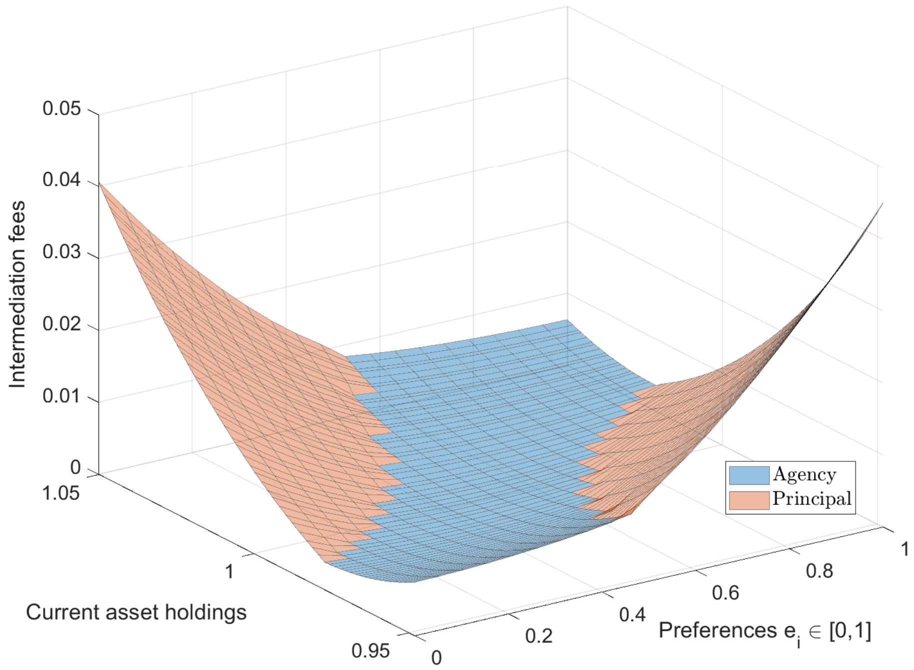

Figure 3. Intermediation fees under each trading mechanism. This figure depicts the fee paid by each customer, conditional on her preference type and current holdings. Orange shaded area refers to principal fees. Blue shaded area refers to (present valued) agency fees.

Overall, the broad features of intermediation costs in LR09 still hold. For example, marginal fees are increasing in the traded volume. Risk aversion implies that, given certain optimal position the marginal trading surplus is increasing in the volume traded. The bargain protocol used imply that fees are linear functions of such surpluses, thus inherit such property [18]. On top of this, two interesting properties regarding the trading mechanism distinction are observed. Firstly, principal fees are on average larger than agency fees. On the one hand, principal traders exchange larger quantities, and thus obtain larger trade surpluses. On the other hand, even conditioning on net trade surplus (surplus after paying fees), principal fees 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 [19]. Secondly principal fees increase at a higher rate when moving

[18] Pinter, Wang and Zou (2021) 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 have both customers and dealers identity. In line with the model here developed, they show that, conditional on the customer identity, trading costs are increasing in trade size.

[19] 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 fees are convex combinations of customers' trade surpluses and dealer costs, I get that, at the thresholds, principal fees exceed

both towards extreme preferences and towards larger trading quantities. When customers trade on agency, they are subject to preference shocks. This imply 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, 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 fees, are relatively flatter [20].

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

4 Liquidity Measures

Recent empirical literature on the US corporate bond market argue that liquidity conditions have changed during the last decade. Particularly, researchers document a shift in trading volume, from immediate principal towards delayed agency trades, accompanied by an increase in immediacy costs (Bessembinder, Jacobsen and Venkataraman, 2018; Schultz, 2017; Choi, Huh and Shin, 2021; Bao, O'Hara, and Zhou, 2018; Anderson and Stulz, 2017; Dick-Nielsen and Rossi, 2019; O'Hara and Zhou, 2021; Kargar et.al., 2021). 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 for each mechanism. Secondly I perform a transaction costs decomposition which accounts for composition effects. These objects are used to analyze 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

(present valued) agency fees exactly by the inventory costs amount.

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

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

[20] Figure B.1 of the Appendix presents fees per dollar traded. All the features mentioned about fees hold if this alternative specification is considered.

preference i trade on principal, on agency, or don't 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 asset supply. The supply of assets is normalized at $A = 1$, so I only need to compute the volume. Principal trades are performed by customers who are waiting to contact a dealer and rather immediate trades, i.e. customers in state $\eta_{[i,a,\omega_1]}$, where $a \in P_i$. These contacts happen at rate α, and the volume traded in each transactions 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 $\eta_{[i,a,\omega_2]}$, 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 is:

$$\mathcal{T}^P = \alpha \sum_{i \in \mathcal{I}} \sum_{a \in P_i} n_{[a,i,\omega_1]} |a_i^P - a|, \tag{17}$$

$$\mathcal{T}^A = \beta \sum_{i \in \mathcal{I}} \sum_{a \in \mathcal{A}^*} n_{[a,i,\omega_2]} |a_i^A - a|. \tag{18}$$

And 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 firstly compute the transaction cost per (*numeraire*) dollar traded, i.e. the ratio of the fee and dollar valued volume 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, fees are arranged at contact with dealers and optimal allocations are chosen at execution. While waiting for execution, customers can suffer preference shocks. Hence, two identical customers contracting the same agency fee might end up trading different volumes. Hence, I need to compute the aggregated realized volume for each contract, $rav_{a,i}$. 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 is:

$$\mathcal{S}^P = \sum_{i \in \mathcal{I}} \sum_{a \in P_i} \frac{n_{[a,i,\omega_1]} |a_i^P - a|}{\mathcal{T}^P} \frac{\phi_{a,i}^P}{|a_i^P - a| p}, \tag{19}$$

$$\mathcal{S}^A = \sum_{i \in \mathcal{I}} \sum_{a \in A_i} \frac{n_{[a,i,\omega_1]} rav_{a,i}}{\mathcal{T}^A} \frac{\phi_{a,i}^A}{rav_{[a,i]} p}. \tag{20}$$

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[21]:

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

The average transaction cost unconditional on trading mechanism is just the weighted average of the 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 fees associated to each transaction and the steady state mass of customer 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 propose a method to decompose the samples over which transaction costs are measured. The idea is to build subsamples according to the trading mechanism choice of customers across different parametrizations. To do this, 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 paramerizations, denoted by q, and compute their steady state trading mechanism sets. Secondly, for each preference type, compute the intersections across parametrizations between these trading mechanisms sets. To ease the exposition, I only consider two parametrizations, $q \in \{0, 1\}$, but the method can be easily extended to account for $Q > 1$ number of parametrizations. Table 2 presents the resulting subsets. Diagonal cells include customers that choose the same trading mechanism under the two scenarios. Conversely, non-diagonal cells includes customers that change their optimal mechanism when facing different scenarios. 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 $q = 1$ [22].

Table 2: Sample decomposition

	P_i^1	A_i^1	NT_i^1
P_i^0	$[P^0, P^1]_i$	$[P^0, A^1]_i$	$[P^0, NT^1]_i$
A_i^0	$[A^0, P^1]_i$	$[A^0, A^1]_i$	$[A^0, NT^1]_i$
NT_i^0	$[NT^0, P^1]_i$	$[NT^0, A^1]_i$	$[NT^0, NT^1]_i$

[21] 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.

[22] If Q 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 $a \in \cap_{q=1}^{Q} P_i^q$.

These subsets allow to define subsamples over which to compute intermediation costs. To this end, I add new notation. Superscripts attached to costs measures indicate both the trading mechanism and the 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 intermediation costs paid under scenario $q = 0$ by customers who trade on principal both under $q = 0$ and $q = 1$. These subsample specific intermediation costs can be properly weighted to generate a decomposition of the overall measure[23]. Consider for example principal costs:

$$\begin{aligned}
\mathcal{S}^{P,0} &= \mathcal{S}^{P,0}_{P^0,P^1} \times w^{P,0}_{P^0,P^1} + \mathcal{S}^{P,0}_{P^0,A^1} \times w^{P,0}_{P^0,A^1} + \mathcal{S}^{P,0}_{P^0,NT^1} \times w^{P,0}_{P^0,NT^1},\\
\mathcal{S}^{P,1} &= \mathcal{S}^{P,1}_{P^0,P^1} \times w^{P,1}_{P^0,P^1} + \mathcal{S}^{P,1}_{A^0,P^1} \times w^{P,1}_{A^0,P^1} + \mathcal{S}^{P,1}_{NT^0,P^1} \times w^{P,1}_{NT^0,P^1},
\end{aligned}$$

$$\begin{aligned}
\text{where} \qquad \mathcal{S}^{P,q}_{\chi} &= \sum_{i\in\mathcal{I}} \sum_{a\in\chi_i} \frac{n^q_{[a,i,\omega_1]}|a_i^{P,q} - a|}{\sum_{i\in\mathcal{I}}\sum_{a\in\chi_i} n^q_{[a,i,\omega_1]}|a_i^{P,q} - a|} \frac{\phi^{P,q}_{a,i}}{|a_i^{P,q} - a|p^q},\\
w^{P,q}_{\chi} &= \frac{\sum_{i\in\mathcal{I}}\sum_{a\in\chi_i} n^q_{[a,i,\omega_1]}|a_i^{P,q} - a|}{\mathcal{T}^{P,q}},
\end{aligned}$$

and χ denotes any subset in Table 2. Similarly, agency costs can be decomposed as:

$$\begin{aligned}
\mathcal{S}^{A,0} &= \mathcal{S}^{A,0}_{A^0,P^1} \times w^{A,0}_{A^0,P^1} + \mathcal{S}^{A,0}_{A^0,A^1} \times w^{A,0}_{A^0,A^1} + \mathcal{S}^{A,0}_{A^0,NT^1} \times w^{A,0}_{A^0,NT^1},\\
\mathcal{S}^{A,1} &= \mathcal{S}^{A,1}_{P^0,A^1} \times w^{A,1}_{P^0,A^1} + \mathcal{S}^{A,1}_{A^0,A^1} \times w^{A,1}_{A^0,A^1} + \mathcal{S}^{A,1}_{NT^0,A^1} \times w^{A,1}_{NT^0,A^1},
\end{aligned}$$

$$\begin{aligned}
\text{where} \qquad \mathcal{S}^{A,q}_{\chi} &= \sum_{i\in\mathcal{I}} \sum_{a\in\chi_i} \frac{n^q_{[a,i,\omega_1]} rav^q_{a,i}}{\sum_{i\in\mathcal{I}}\sum_{a\in\chi_i} n^q_{[a,i,\omega_1]} rav^q_{a,i}} \frac{\phi^{A,q}_{a,i}}{rav^q_{a,i}p^q},\\
w^{A,q}_{\chi} &= \frac{\sum_{i\in\mathcal{I}}\sum_{a\in\chi_i} n^q_{[a,i,\omega_1]} rav^q_{a,i}}{\mathcal{T}^{A,q}},\\
rav^{\chi}_{a,i} &= (1-\hat{\delta})|a_i^{A,\chi} - a| + \hat{\delta}\sum_{j\in\mathcal{I}} \pi_j |a_j^{A,\chi} - a|.
\end{aligned}$$

As can be seen, transaction costs are the weighted sum of the costs paid by customers who, facing a parametric change, would react differently regarding their trading mechanism choice. Particularly, some customers will decide to continue performing their trades using the same mechanism, and some others will decide to migrate whether towards another type of trade or not to trade at all and wait for the next contact with a dealer. I call these customers non-migrant and migrant, respectively. Finally, I can decompose the change in intermediation costs for each mechanism due to a parametric change. Consider $q = 0$ as the initial

[23] See Appendix A.5. for details.

scenario, and $q = 1$ as the new one.

$$\begin{aligned}\Delta\mathcal{S}^{P} = \mathcal{S}^{P,1} - \mathcal{S}^{P,0} = {} & \underbrace{\mathcal{S}^{P,1}_{P^0,P^1} \times w^{P,1}_{P^0,P^1} - \mathcal{S}^{P,0}_{P^0,P^1} \times w^{P,0}_{P^0,P^1}}_{\text{Principal non-migrants}} \\ & + \underbrace{\mathcal{S}^{P,1}_{A^0,P^1} \times w^{P,1}_{A^0,P^1} + \mathcal{S}^{P,1}_{NT^0,P^1} \times w^{P,1}_{NT^0,P^1}}_{\text{Inflow migration}} \\ & - \underbrace{\mathcal{S}^{P,0}_{P^0,A^1} \times w^{P,0}_{P^0,A^1} - \mathcal{S}^{P,0}_{P^0,NT^1} \times w^{P,0}_{P^0,NT^1}}_{\text{Outflow migration}}, \end{aligned} \tag{21}$$

$$\begin{aligned}\Delta\mathcal{S}^{A} = \mathcal{S}^{A,1} - \mathcal{S}^{A,0} = {} & \underbrace{\mathcal{S}^{A,1}_{A^0,A^1} \times w^{A,1}_{A^0,A^1} - \mathcal{S}^{A,0}_{A^0,A^1} \times w^{A,0}_{A^0,A^1}}_{\text{Agency non-migrants}} \\ & + \underbrace{\mathcal{S}^{A,1}_{P^0,A^1} \times w^{A,1}_{P^0,A^1} + \mathcal{S}^{A,1}_{NT^0,A^1} \times w^{A,1}_{NT^0,A^1}}_{\text{Inflow migration}} \\ & - \underbrace{\mathcal{S}^{A,0}_{A^0,P^1} \times w^{A,0}_{A^0,P^1} - \mathcal{S}^{A,0}_{A^0,NT^1} \times w^{A,0}_{A^0,NT^1}}_{\text{Outflow migration}}. \end{aligned} \tag{22}$$

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 that 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.

Note that equations (21) and (22) provide a natural way of defining measures of intermediation costs free of composition effects. If the samples within trading mechanism were held constant, non-migrant customers would have full weight in all scenarios. Therefore, I define the composition-free measures of intermediation 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 intermediation 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, are defined as the fraction of the change in intermediation costs due to migration.

$$\tilde{\mathcal{S}}^P(q) \equiv \mathcal{S}^{P,q}_{P^0,P^1}, \tag{23}$$

$$\tilde{\mathcal{S}}^A(q) \equiv \mathcal{S}^{Aq}_{A^0,A^1}, \tag{24}$$

$$\Delta\tilde{\mathcal{S}}^P \equiv \mathcal{S}^{P,1}_{P^0,P^1} - \mathcal{S}^{P,0}_{P^0,P^1}, \tag{25}$$

$$\Delta\tilde{\mathcal{S}}^A \equiv \mathcal{S}^{A,1}_{A^0,A^1} - \mathcal{S}^{A,0}_{A^0,A^1}, \tag{26}$$

$$CE^P \equiv 1 - \Delta\tilde{\mathcal{S}}^P/\Delta\mathcal{S}^P, \tag{27}$$

$$CE^A \equiv 1 - \Delta\tilde{\mathcal{S}}^A/\Delta\mathcal{S}^A. \tag{28}$$

The introduction of composition-free measures of intermediation costs change sheds light over the necessary conditions for the existence of composition effects mentioned in the introduction of this paper. In first place, the absence of migrating trades 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 in each mechanism. Finally, as long as the difference between costs paid by migrants and non-migrants are driven by unobservable characteristics, empirical estimates would include a composition effect bias. Our model suggest that such 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 Numerical Exercises

In this section I use the model to revisit the evidence related to the two major changes observed in the US corporate bond markets in the last decade. Firstly, I address the introduction of post 2008 financial crisis regulations by increasing the models' inventory costs. Secondly, 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 across 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.

5.1 Increase in Inventory Costs

Inventory costs are captured by θ, the marginal cost per dollar traded on principal. This parameter was set at $0.1b.p.$ taking as reference the leverage ratio requirement (LRR), which is arguably the most tightly binding constraint for bank-dealers in the US (Greenwood et. al. 2017, Duffie 2018). Post financial crisis regulation requires a ratio of capital to non-risk-weighted assets of no less than 5% for global systemically important banks. Therefore, in order to match such figure, in this exercise I consider $\theta = 0.7b.p.$. Figure 4 shows the policy functions resulting from the new calibration.

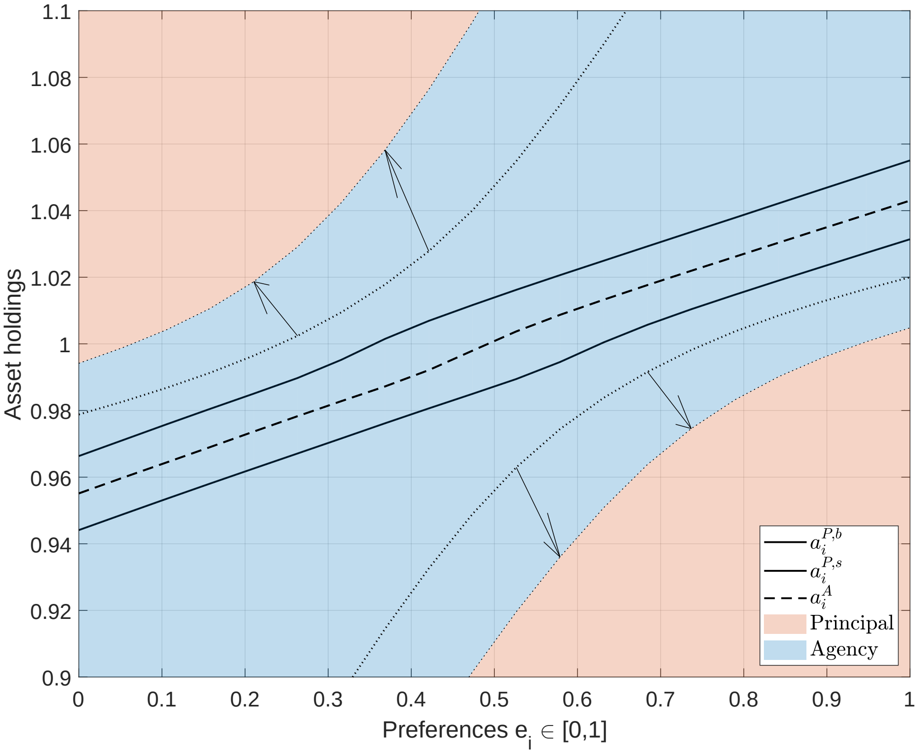

Figure 4. Inventory Costs Increase. Optimal asset holdings and trade choice under $\theta = 0.7b.p.$. This figure depicts the policy functions of each customer, conditional on her preference type and current holdings. The lower and upper solid lines represent the buyer's and seller's optimal allocations under the principal trade, $a^{P,b}$ and $a^{P,s}$, respectively. The dashed line represents the optimal allocation under agency trading, a^A. Regarding the trading mechanism choice, the principal and agency regions are shaded in orange and blue, respectively. To ease the comparison across calibrations, the trading mechanism thresholds of the baseline calibration are depicted as dotted lines within the agency region, and the arrows denote its expansion.

An increase in dealers' inventory costs makes immediate principal trades more expensive. Larger inventory costs are translated to customers through higher fees, thus effective prices of principal trading are larger when customers buy and smaller when they sell. As a consequence, the optimal volume traded on principal for each preference type is reduced. This can be observed by buyers having lower optimal allocations and sellers having higher optimal allocations. The corollary being that the region in which a customer principal trading surplus is not enough to cover inventory costs, $a \in [a_i^{P,b}, a_i^{P,s}]$, widens. However, as in the baseline calibration, all those customers decide to trade on an agency basis rather than waiting without trading for another contact with dealers. To highlight the migration of trades across mechanisms, Figure 4 includes the baseline calibration threshold as dotted lines found within the current calibration agency region. As can be seen, an increase in inventory costs makes customers to migrate away from principal towards agency trading.

Figure 5 presents the liquidity measures computed for $\theta \in [0.1b.p., 0.7b.p.]$. 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, less 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 makes the average volume per trade to decrease in such mechanism and in the entire sample. 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.

Intermediation 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 fees, and so principal trading costs mechanically rises. Comparing the two extremes of the inventory costs range considered, average principal cost increases 55.8 bp, which represent a percentage change of 38.9%. Despite of non being directly related to inventory costs, agency costs increase as well by 7.8 bp, representing a smaller percentage change of 10.9%. The effect of inventory costs over agency costs is due to a general equilibrium effect in which the value function steepens in asset holdings. When immediate reallocation becomes more expensive, every position is expected to be held longer, thus the difference between the lifetime utility of holding two different positions widens and the trade surplus increases.

The correlations between inventory costs, migration across mechanisms and average intermediation 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 costs measures has been largely overlooked. Panel B of Figure 5 accounts for such composition effects using the proposed decomposition. I use dashed lines to plot the composition-free measures, $\tilde{\mathcal{S}}^P$ and $\tilde{\mathcal{S}}^A$, for each trading mechanism. The comparison of average and composition-free measures allows us to gauge the size of the bias.

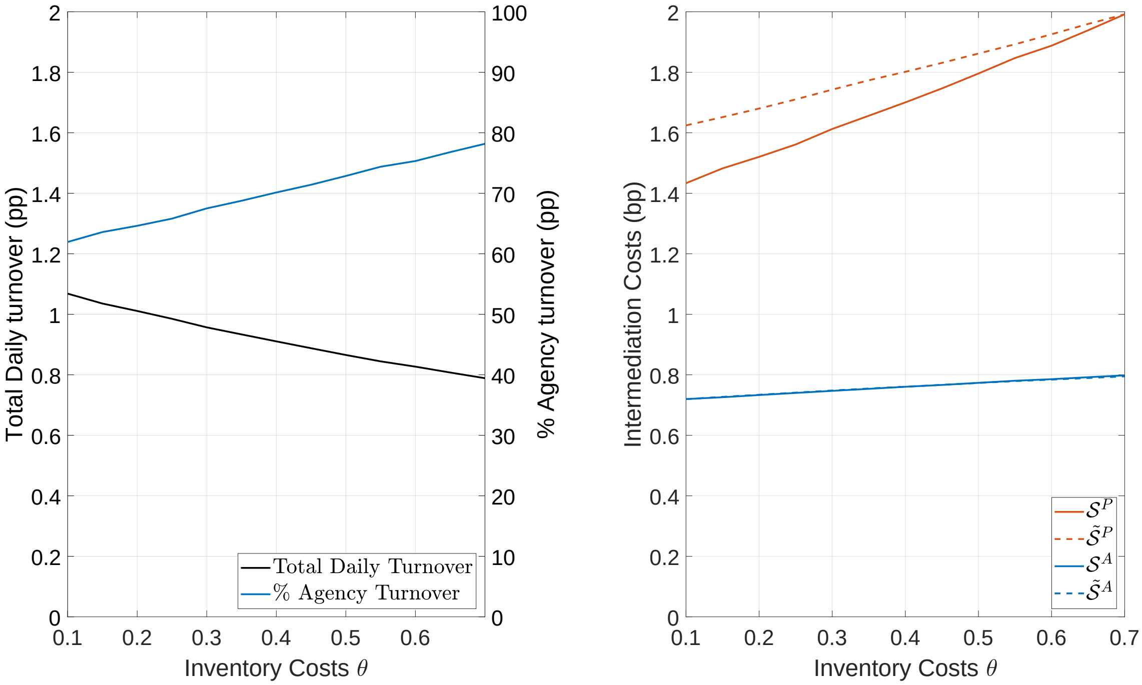

Figure 5. Liquidity Measures when Inventory Costs Increase. Panel A (left) presents the steady state total daily turnover rate, $\mathcal{T}$, and the agency percentage of such figure, $\mathcal{T}^{\mathcal{A}}/\mathcal{T}$, across $\theta \in [0.1b.p., 0.7b.p.]$. Panel B (right) presents the steady state weighted average intermediation costs for both mechanisms across $\theta \in [0.1b.p., 0.7b.p.]$. Solid lines represent the average intermediation costs, $\mathcal{S}^{\mathcal{P}}$ and $\mathcal{S}^{\mathcal{A}}$, and dashed lines represent the composition-free costs, $\tilde{\mathcal{S}}^{\mathcal{P}}$ and $\tilde{\mathcal{S}}^{\mathcal{A}}$.

Let me start by addressing principal costs. The migration pattern presented in Figure 4 tell us that principal customers can be split into non-migrants and outflowing migrants. When marginal inventory costs are set at $\theta = 0.1b.p.$, the composition-free measure, i.e. the intermediation cost payed by non-migrants, is 13.3% larger than the mechanism's average. Such difference is understood going back to Figure 4, where it is observed that non-migrant principals are customers with relatively more extreme preferences and more extreme asset positions, both characteristics associated with larger intermediation payments. As inventory costs increase, some customers migrate towards agency trading and the proportion of non-migrants increase. 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 intermediation costs is smaller than that of the mechanism's average. Particularly, $\Delta\tilde{\mathcal{S}}^P = 36.7$bp, 19.1bp less than the average figure. This difference is explained by the composition effect. When $\Delta\mathcal{S}^P$ is computed, it includes the baseline sample of customers who would migrate away from the mechanism, and thus would not be present in the final comparison sample. Given that those out-flowing migrants pay less than the average, $\Delta\mathcal{S}^P$ is biased upwards. Such composition effect bias accounts for

$CE^P = 34.2\%$ of the average costs increase[24]. In other words, when inventory costs increase, the average willingness to pay of the resulting sample increases, given that those customers that remain trading on principal are the ones who had higher willingness to pay before costs increased. Therefore, the average costs change captures this average willingness to pay increase, and is consequently biased upwards.

Regarding agency trades, the migration pattern associated with increasing inventory costs tell us that customers in this mechanism can be separated into non-migrants and inflowing migrants. Given that no agency customer will migrate away when increasing inventory costs, at $\theta = 0.1 b.p.$ 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 0.4% smaller costs than the mechanisms' average. This mild difference contrasts with the principal case, and its explained by the small fee 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 intermediations costs measures. Matter of fact, when comparing the two extremes of the parametric range considered, the composition-free measure equals $\Delta\tilde{\mathcal{S}}^A = 7.5$bp, only 0.3bp below $\Delta\mathcal{S}^A$. Correspondingly, for the agency case I find a mild composition effect bias of $CE^A = 4.1\%$.

To sum up, the model's predictions are in line with both the empirical and the theoretical literature that studies the effects of rising the intermediaries' inventory costs. In a nutshell, the provision of immediacy services becomes more expensive, and intermediation shifts away from inventory related towards inventory free mechanisms, i.e. principal and agency trades, respectively. Nevertheless, the model also suggest that transaction costs measures should be revisited, considering the impact that composition effects may have on them. Particularly, I find that these effects accounts for around a third of the increase in principal costs, and for a smaller figure, around 4%, or agency costs increases.

5.2 Decrease in the Execution Delay

The increase of electronic trading (in contrast with voice trading) made easy for dealers to match counterparties in agency trades (O'Hara and Zhou (2021)). I model this market change as a reduction in the execution delay of such mechanism [25]. Such delay is captured in the model by β. In the baseline calibration, this parameter what set so that customers wait on expectation three days for their trades to be executed. I use the model to analyze the impact of decreasing such delay to one day. This new calibration implies that contacting a dealer takes the same time as to execute agency trades. The new policy function are presented

[24] Whenever $\tilde{\mathcal{S}}^P$ and $\mathcal{S}^P$ are linear on θ, the composition effect bias is constant. Figure 5 indicate that the computed slopes can be well approximated by linear functions.

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

in Figure 6.

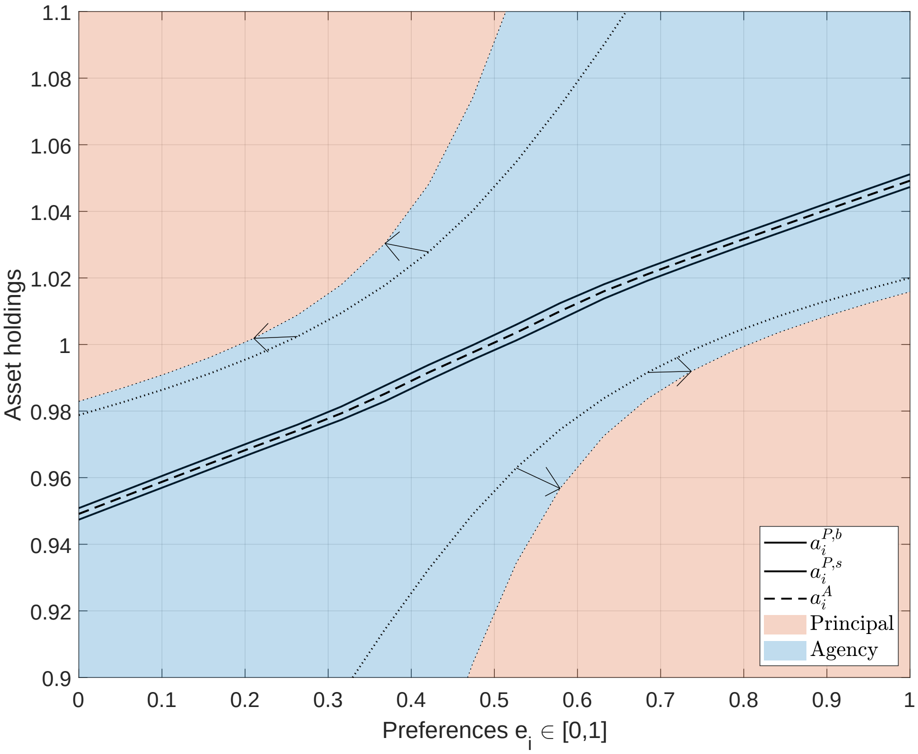

Figure 6. Execution Delays Decrease. Optimal asset holdings and trade choice under $\beta = 1$. This figure depicts the policy functions of each customer, conditional on her preference type and current holdings. The lower and upper solid lines represent the buyer's and seller's optimal allocations under the principal trade, $a^{P,b}$ and $a^{P,s}$, respectively. The dashed line represents the optimal allocation under agency trading, a^A. Regarding the trading mechanism choice, the principal and agency regions are shaded in orange and blue, respectively. To ease the comparison across calibrations, the trading mechanism thresholds of the baseline calibration are depicted as dotted lines within the agency region, and the arrows denote its expansion.

Note firstly that execution delays plays no important role when customers choose their optimal asset positions. The case is obvious regarding principal trades, but is more subtle in the agency case. As pointed out by equation (4), when customers choose their optimal agency positions they do it at execution. Therefore the waiting time does not play any role in such choice. Consequently, the optimal asset positions in the baseline and in the new calibration do not depart significantly. Notwithstanding these similarities, a reduction in execution costs affects the trading mechanism choice. Smaller execution delays implies that agency customers need to hold unwanted positions for less time, thus the relative attractiveness of such contract increase. Consequently, customers with preference type - asset positions close to the baseline calibraition thresholds migrate away from principal towards agency, being the effect stronger for customers with preference types closer to the mean.

The liquidity measures computed for the range $\beta \in [\frac{1}{3}, 1]$ are presented in Figure 7. Panel A presents the daily turnover as well as the percentage explained by agency trades. Increasing the execution speed of non-immediate contracts has both a direct and an indirect effect over turnover figures. The direct effect is to increase the extensive margin of both agency and principal trades. On the one hand, the amount 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 on a daily basis and optimally chose whether to arrange new principal or new agency contracts. The indirect effect is that the intensive margins are also impacted. This happens in despite of the optimal allocations not being significantly modified. Firstly, the migrating customers make the average volume traded in both principal and agency contract to be larger. Figure 6 shows, for each preference type, the expansion of both the maximum and the minimum distances between the current and the optimal position under agency and principal trades, respectively. Secondly, a faster execution implies that agency customers are more likely to avoid a preference shock while waiting for 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 allocations, trading according to the current type implies a decrease in the average agency volume per trade [26]. Overall, these effects jointly explain a 53.6% increase in the daily turnover and a slight 4.4% decrease in the agency share.

Panel B of Figure 7 shows the intermediation 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 $\Delta\mathcal{S}^P =$ 5.8bp and agency costs rise by $\Delta\mathcal{S}^A = 20$bp. 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 over the implied trading surplus of each customer. Therefore, keeping samples constant, principal costs should not significantly change. Accordingly, composition-free principal costs only have a slight increase of $\Delta\tilde{\mathcal{S}}^P = 0.1$bp and almost the entire increase in average principal costs are due to composition effects, $CE^P = 98.7\%$. The explanation is found in Figure 6. Principal customers with relatively moderate preferences and asset positions, characteristics associated with low intermediation payments, migrate away from the mechanism, increasing the average willingness to pay of the remaining sample. Regarding agency trades, a reduction in expected delays has as direct positive impact on the expected trade surplus of every agency customer: unwanted positions can be optimally exchanged faster. I compute an increase in the agency composition-free costs of $\Delta\tilde{\mathcal{S}}^A = 20.3$bp. 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

[26] 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.

negligible, computed at $CE^A = -1.5\%$.

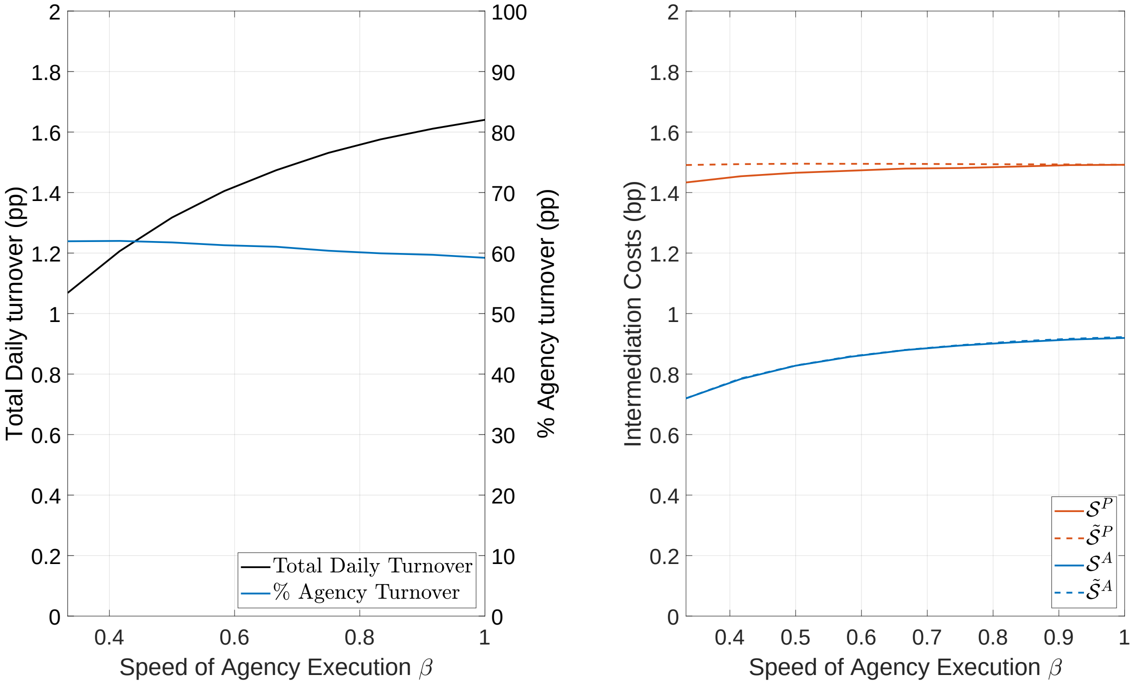

Figure 7. Liquidity Measures when Execution Delays Decrease. Panel A (left) presents the steady state total daily turnover rate, $\mathcal{T}$, and the agency percentage of such figure, $\mathcal{T}^{\mathcal{A}}/\mathcal{T}$, across $\beta \in [1/3, 1]$. Panel B (right) presents the steady state weighted average intermediation costs for both mechanisms across $\beta \in [1/3, 1]$. Solid lines represent the average intermediation costs, $\mathcal{S}^{\mathcal{P}}$ and $\mathcal{S}^{\mathcal{A}}$, and dashed lines represent the composition-free costs, $\tilde{\mathcal{S}}^{\mathcal{P}}$ and $\tilde{\mathcal{S}}^{\mathcal{A}}$.

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 rising the transaction costs of such mechanism. This result complements the supply shifts across mechanisms due to the decrease in search and match costs documented by the empirical literature. 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.

6 Conclusion

Researchers have documented a recent change in the liquidity profile of OTC markets due to new regulations and technology improvements. Both the type of intermediation provided by dealers, i.e. principal or agency trades, and the cost of each of them were affected. For example, the consensus points that the rise in inventory cots due to the post 2008 financial crisis regulations increased the costs of immediate principal trades, shifting volume towards agency intermediation. This paper revisit such evidence, by arguing that the joint determination of the terms of trade and trading mechanisms in such markets implies that intermediation cost measures can be biased by composition effects. Whenever market conditions change, customers may optimally migrate from one mechanism to another. If migrating and non-migrating customers pay different transaction costs, the estimates of costs' changes conditional on trading mechanism will be biased.

To account for such bias, I develop a search model with heterogeneous customers, where two trading mechanisms are optimally chosen. Principal trading is immediate but expensive, while agency trading is slow but cheaper. This trade-off defines each customers' optimal trading mechanisms and the average liquidity measures therein. The model allows to decompose the distribution of transaction costs into samples where composition effects can be controlled for. I calibrate the model to the US corporate bond market and perform numerical exercises motivated by recent developments in such market. In those exercises, a fraction of principal customers migrate towards agency. Given that those customers that did not migrate paid on average higher fees, the change in average costs is upward biased. Particularly, composition effects account for a third of the change in immediate 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 contributes to the debate of whether stricter financial regulations are welfare improving. If immediate intermediation costs have not increased as much as it was previously thought, new regulations may have improved financial soundness at a lower expense.

References

An, Y. (2020). *"Competing with inventory in dealership markets"*, Working Paper, Johns Hopkins University.

An, Y., and Zheng, Z. (2020). *"Conflicted immediacy provision"*, Paper, Johns Hopkins University and UC Berkeley.

Anderson, M. and Stulz, R. M. (2017). *"Is Post-Crisis Bond Liquidity Lower?"* NBER Working Papers 23317.

Bao, J., O'Hara, M. and Zhou, X. (2018). *"The Volcker Rule and corporate bond market making in times of stress"*, Journal of Financial Economics 130, 95-113.

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

Choi, J., Huh, Y. and Shin S. (2021). *"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.

Cohen, A., Kargar, M., Lester, B and Weill, P-O (2021). *Inventory, Market Making, and Liquidity: Theory and Application to the Corporate Bond Market* working paper, Conferences: SaMMF, SED 2021.

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

Duffie, D. (2012). *"Market making under the proposed Volcker Rule"*, Rock Center for Corporate Governance at Stanford University Working Paper No. 106.

Duffie, D. (2018). *"Post Crisis Bank Regulations and Financial Market Liquidity"*, Baffi Lecture, Banca d'Italia, Rome, Italy.

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

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

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

Hendershott, T. and Madhavan, A. (2015). *“Click or call? Auction versus search in the over-the-counter market”*. Journal of Finance 70, 419–447.

Hendershott, T., Li, D., Livdan, D., and Schürhoff, N. (2020). *“True Cost of Immediacy”*. Swiss Finance Institute Research Paper No. 20-71.

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

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

Kirkby, R. (2017). *”Convergence of Discretized Value Function Iteration*, Computational Economics 49, 117-153.

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

Miao, J. (2006). *“A search model of centralized and decentralized trade”*. Review of Economic Dynamics 9, 68-92.

O’Hara M. and Zhou X. (2021). *“The electronic evolution of corporate bond dealers”*, Journal of Financial Economics 140, 368-390.

Pinter, G., Wang, C. and Zou, J. (2021). *“Size discount and size penalty: trading costs in bond markets”*. Jacobs Levy Equity Management Center for Quantitative Financial Research Paper.

Saar, G., Sun J., Yang, R. and Zhu, H. (2020). *“From market making to matchmaking: Does bank regulation harm market liquidity?”*, working paper, Cornell University, Harvard Business School, and MIT.

Schultz, P. (2017). *“Inventory management by corporate bond dealers”*, working paper, University of Notre Dame.

Shen, J. (2015) *“Exchange or OTC market: a search-based model of market fragmentation and liquidity”*, working paper.

Stokey, N., Lucas, R. E. and Prescott, E. C. (1989). *“Recursive Methods in Economic Dynamics”*, Cambridge, Harvard University Press.

Wu, B. (2022). *"Post-Crisis Regulations, Trading Delays, and Increasing Corporate Bond Liquidity Premium"*, NYU Stern School of Business.

Appendix A

A.1 Bargaining Outcomes

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

$$[a_i^P(t), \phi_i^P(a,t)] = \arg\max_{(a',\phi')} \Big\{ V_i(a',t) - V_i(a,t) - p_t[a'-a] - \phi' \Big\}^{1-\eta} \Big\{ \phi' - f(a'-a) \Big\}^{\eta}$$

$$= \arg\max_{(a',\phi')} (1-\eta)\ln[\underbrace{V_i(a',t) - V_i(a,t) - p_t[a'-a] - \phi'}_{A}] + \eta\ln[\underbrace{\phi' - f(a'-a)}_{B}].$$

$$\begin{aligned} \text{FOC}_{\phi'}: \quad & -[1-\eta]A^{-1} + \eta B^{-1} = 0 \qquad \text{(assume interior solution)} \\ & \eta A - [1-\eta]B = 0 \\ & \eta[V_i(a',t) - V_i(a,t) - p_t[a'-a]] + [1-\eta]f(a'-a) = \phi_i^P(a,t) \end{aligned}$$

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

$$\begin{aligned} a_i^P(a,t) = \arg\max_{a'} & \Big\{ [1-\eta]\big[V_i(a',t) - V_i(a,t) - p_t[a'-a] - f(a'-a)\big] \Big\}^{1-\eta} \\ & \Big\{ \eta\big[V_i(a',t) - V_i(a,t) - p_t[a'-a] - f(a'-a)\big] \Big\}^{\eta} \\ \arg\max_{a'} \quad & V_i(a',t) - V_i(a,t) - p_t[a'-a] - f(a'-a). \end{aligned}$$

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 bargain power. In other words, investor's utility flow is equal when trading at α rate with a dealer with η bargain power, than trading at a slower rate $[1-\eta]\alpha$ with a dealer with no bargain power. Let's replace the optimal terms of trade from equations (3), (4), (5) and (6) into equation (1).

$$
\begin{aligned}
V_i(a,t) = \mathbb{E}_{i,t}\Big[& \int_t^{T_\alpha} e^{-r[s-t]} u_{k(s)}(a) ds \\
& + e^{-r[T_\alpha - t]} \max \Big\{ [1-\eta]\big[V_{k(T_\alpha)}(a^P_{k(T_\alpha)}, T_\alpha) - p_{T_\alpha}[a^P_{k(T_\alpha)} - a] - f(a^P_{k(T_\alpha)} - a)\big] + \eta V_i(a,t), \\
& [1-\eta]\Big[\int_{T_\alpha}^{T_\beta} e^{-r[s-T_\alpha]} u_{k(s)}(a) ds + e^{-r[T_\beta - T_\alpha]} \big[V_{k(T_\beta)}(a^A_{k(T_\beta)}, T_\beta) - p_{T_\beta}[a^A_{k(T_\beta)} - a]\big]\Big] + \eta V_i(a,t) \Big\}\Big].
\end{aligned}
$$

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 goes to zero:

$$
\begin{aligned}
V_i(a,t) = & \frac{1}{1+rh}\Big[u_i(a)h + Pr[\tau = \tau_\delta \leq h]\Big[\sum_j \pi_j V_j(a, t+h)\Big] \\
& + Pr[\tau = \tau_\alpha \leq h]\Big[[1-\eta] \max \Big\{ V_i(a^P_i, t+h) - p_{t+h}[a^P_i - a] - f(a^P_i - a), \\
& , \mathbb{E}_{i,t+h}\Big[\int_{t+h}^{T_\beta} e^{-r[s-(t+h)]} u_{k(s)}(a) ds + e^{-r[T_\beta - (t+h)]} \big[V_{k(T_\beta)}(a^A_{k(T_\beta)}, T_\beta) - p_{T_\beta}[a^A_{k(T_\beta)} - a]\big]\Big]\Big\} + \eta V_i(a, t+h)\Big] \\
& + Pr[\tau > h] V_i(a, t+h)\Big] \\
& = \frac{1}{1+rh}\Big[u_i(a)h + \delta h\Big[\sum_j \pi_j V_j(a, t+h)\Big] + \alpha h \Big[[1-\eta] \max \Big\{ V_i(a^P_i, t+h) - p_{t+h}[a^P_i - a] - f(a^P_i - a), \\
& \mathbb{E}_{i,t+h}\Big[\int_{t+h}^{T_\beta} e^{-r[s-(t+h)]} u_{k(s)}(a) ds + e^{-r[T_\beta - (t+h)]} \big[V_{k(T_\beta)}[a^A_{k(T_\beta)}, T_\beta] - p_{T_\beta}[a^A_{k(T_\beta)} - a]\big]\Big]\Big\} + \eta V_i(a, t+h)\Big] \\
& + [1 - \delta h - \alpha h] V_i(a, t+h)\Big] \\
& = \frac{1}{1+rh}\Big[u_i(a)h + \delta h\Big[\sum_j \pi_j V_j(a, t+h)\Big] + \underbrace{\alpha[1-\eta]h}_{Pr[\tau' = \tau_\kappa \leq h]} \Big[\max \Big\{ V_i(a^P_i, t+h) - p_{t+h}[a^P_i - a] - f(a^P_i - a), \\
& \mathbb{E}_{i,t+h}\Big[\int_{t+h}^{T_\beta} e^{-r[s-(t+h)]} u_{k(s)}(a) ds + e^{-r[T_\beta - (t+h)]} \big[V_{k(T_\beta)}(a^A_{k(T_\beta)}, T_\beta) - p_{T_\beta}[a^A_{k(T_\beta)} - a]\big]\Big]\Big\}\Big] \\
& + \underbrace{[1 - \delta h - \alpha[1-\eta]h]}_{Pr[\tau' > h]} V_i(a, t+h)\Big],
\end{aligned}
$$

where $\tau' = \min\{\tau_\delta, \tau_\kappa\}$ and τ_κ is the bargain-adjusted time it takes to contact a dealer, wich 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, $\eta' = 0$.

$$\begin{aligned}
V_i(a,t) = \mathbb{E}_{i,t}\Big[& \int_t^{T_\kappa} e^{-r[s-t]} u_{k(s)}(a)ds \\
& + e^{-r[T_\kappa - t]} \max\Big\{ V_{k(T_\kappa)}(a^P_{i(T_\kappa)}, T_\kappa) - p_{T_\kappa}[a^P_{i(T_\kappa)} - a] - f(a^P_{i(T_\kappa)} - a), \\
& \int_{T_\kappa}^{T_\beta} e^{-r[s-T_\kappa]} u_{k(s)}(a)ds + e^{-r[T_\beta - T_\kappa]}\big[V_{k(T_\beta)}(a^M_{i(T_\beta)}, T_\beta) - p_{T_\beta}[a^M_{i(T_\beta)} - a]\big]\Big\}\Big].
\end{aligned}$$

Appendix 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 derive in Appendix A.2 over some small time horizon h, and let h goes to zero [27].

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

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

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 allocation chosen at execution (see equation (6)). Note that, while waiting, the customer

[27] For the ease of exposition I removed time subscripts.

might change his preferences, which is reflected in the second term of the right hand side of the above equation. Expression $V_i^A(a)$ can be further manipulated to be written as a function of $V_i(a)$. Let me first get the expression for $\sum_j \pi_j V_j^A(a)$:

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

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 into 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:

$$\begin{aligned}
V_i(a) = \bar{U}_i^\kappa(a) + \hat{\kappa}\Big[&[1-\hat{\delta}]\max\Big\{V_i(a_i^P) - p[a_i^P - a] - f(a_i^P - a), \bar{U}_i^\beta(a) + \hat{\beta}\big[\bar{V}_i^A - p[\bar{a}_i^A - a]\big]\Big\} \\
&+ \hat{\delta}\sum_j \pi_j \max\Big\{V_j(a_j^P) - p[a_j^P - a] - f(a_j^P - a), \bar{U}_j^\beta(a) + \hat{\beta}\big[\bar{V}_j^A - p[\bar{a}_j^A - a]\big]\Big\}\Big], \qquad (29)
\end{aligned}$$

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

Appendix A.4. 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 of Stokey and Lucas (1989) - SL89 -, which requires the functions mapped by T to be continuous and bounded. Define $S = R_+ \times \{1, .., I\}$, $C = \{g : S \to R \mid g(a, i)$ is continuous in a 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[28].

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 satisfy 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 $[T(V + \epsilon)](i, a) \leq$

[28] 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 guarantee by the indifference condition that originates the kinks. See Kirkby (2017) for a proof of the convergence of the computational solution to the true solution using discretized value function iteration.

$[TV](i,a) + \lambda\epsilon$ for all $V \in C$, {a,i}$\in S$ and $\epsilon \geq 0$. Consider $[T(V+\epsilon)](i,a)$:

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

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

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

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

where the last inequality comes from the fact that subtracting an 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 choose 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 $[\hat{\kappa}\hat{\beta}, \hat{\kappa}]$.

Appendix A.5. Intermediation Fees per Dollar Traded

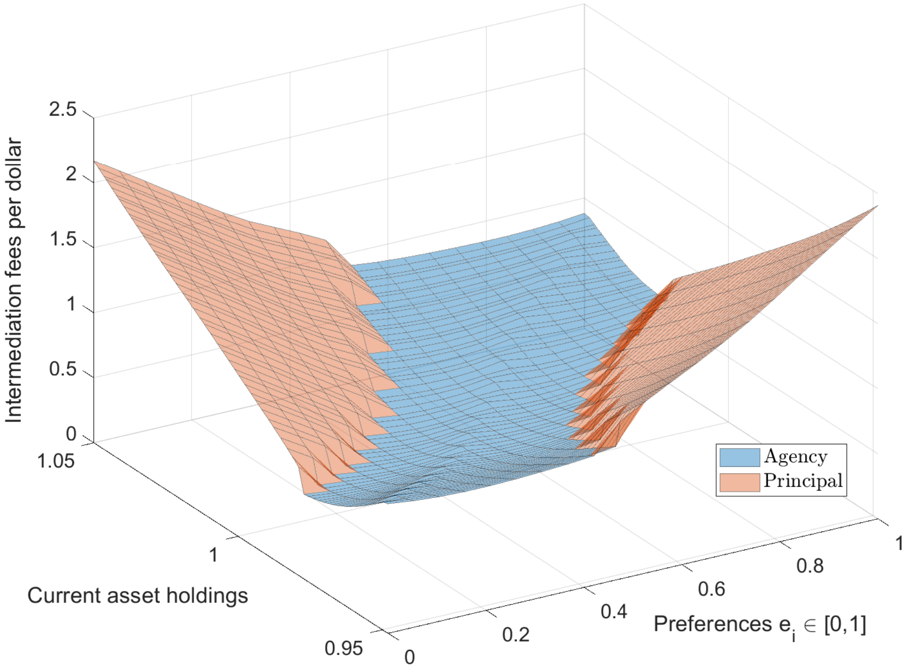

Figure A.1. Intermediation costs per dollar traded under each trading mechanism. This figure depicts the intermediation fees per dollar traded paid by each customer, conditional on her preference type and current holdings, and expressed in basic points. Agency fees are computed using the expected volume traded for each customer, according to the optimal $\bar{a}_i^A$, and expressed in present value at the moment of contact with the dealer.

Appendix A.6. Transaction Costs Decomposition

Here I present the algebra steps needed to decomposed the intermediation costs measures in equations (19) and (20), particularly for the parametrization $q = 0$.

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

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

where

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