Stablecoin Runs and the Centralization of Arbitrage*

Yiming Ma[†] Columbia Business School

Yao Zeng[‡] Wharton

Anthony Lee Zhang[§] Chicago Booth

This version: March 2023

Click **here** for the latest version

*We thank Campbell Harvey, Xavier Giroud, Urban Jermann, Nellie Liang, Anna Pavlova, and Mila Sherman for detailed comments, and conference and seminar participants at AFA AFFECT, Duke/UNC/Milken DeFi Conference, Texas A&M, UT Dallas, and Wharton. Junyi Hu, Liming Ning, Haichuan Wang, Yuming Yang, and Max Yang provided excellent research assistance. All errors are ours.

[†]Finance Division, Columbia Business School, 3022 Broadway, Uris Hall 820, New York, NY 10027. Email: ym2701@gsb.columbia.edu.

[‡]Wharton School, University of Pennsylvania, Steinberg-Dietrich Hall Suite 2300, Philadelphia, PA 19104. Email: yaozeng@wharton.upenn.edu.

[§]Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave. Chicago, IL 60637. Email: anthony.zhang@chicagobooth.edu.

Stablecoin Runs and the Centralization of Arbitrage

Click here for the latest version

Abstract

We analyze the run risk of USD-backed stablecoins and uncover a dilemma between stablecoins' price stability and financial stability. Stablecoin runs bear important financial stability implications through the fire sale of US dollar assets like bank deposits, Treasuries, and corporate bonds. We show that panic runs exist even though general investors only trade stablecoins in secondary markets with flexible prices. Run incentives are reinstated by stablecoin issuers' liquidity transformation and the fixed $1 at which arbitrageurs redeem stablecoins for cash in the primary market. We discover that more efficient arbitrage amplifies run risk. This explains why stablecoin issuers only authorize a small set of arbitragers even though it comes at the expense of maintaining a stable secondary price. In other words, the centralization of arbitrage embeds an inherent tradeoff between run risk and price stability. Our findings are based on a model and a novel dataset on stablecoin redemptions, trading, and reserve assets. Calibrating our model, we find a higher run risk for USDT, the largest stablecoin, compared to USDC, the second-largest stablecoin. However, even USDC bears significant run risk due to its less concentrated arbitrage and more concentrated deposit holdings.

1 Introduction

Stablecoins are blockchain assets whose value is claimed to be stable at $1. The main type of stablecoins, fiat-backed stablecoins, attempt to achieve price stability by promising to back each stablecoin token with at least $1 in US dollar-denominated assets, which range from bank deposits and Treasuries to corporate bonds and loans. The potential for stablecoins to become the safe asset for the blockchain ecosystem as well as a means of payment for real purchases has contributed to their meteoric rise. The six largest US dollar-backed stablecoins have grown from $5.6 billion in asset size at the beginning of 2020 to exceed $130 billion at the beginning of 2022.

The rapid expansion of stablecoins has also raised financial stability concerns.[1] Terra USD, one of the largest algorithmic stablecoins, experienced a sharp run in May 2022, which led to its collapse in a week (Liu, Makarov and Schoar, 2023).[2] In March 2023, Circle's USDC, the second largest fiat-backed stablecoin, also experienced a run amid the collapse of Silicon Valley Bank with its price plummeting by more than 15% within a few hours. These financial stability concerns have also been a major driving force behind efforts to introduce central bank digital currencies (Brunnermeier, James and Landau, 2019, Duffie, 2019, Auer, Frost, Gambacorta, Monnet, Rice and Shin, 2022, Makarov and Schoar, 2022).

Unlike other crypto assets, fiat-backed stablecoins are directly connected to the traditional financial system through their US dollar asset holdings. A run on them would not only lead to losses for stablecoin investors but could also contract bank deposit funding, strain US Treasury markets, and induce the fire sales of illiquid assets like corporate bonds. These ramifications may become even more pronounced going forward as stablecoins potentially become a more widely adopted means of payment and an increasingly important holder of financial assets. Thus, it is essential to understand whether runs could materialize in the future and what design features of stablecoins could affect their occurrence.

[1] For example, see, G7 Working Group and others, 2019, "Investigating the Impact of Global Stablecoins"; ECB, 2020, "Stablecoins: Implications for monetary policy, financial stability, market infrastructure and payments, and banking supervision in the euro area"; BIS, 2020, "Stablecoins: potential, risks and regulation"; and IMF, 2021, "The Crypto Ecosystem and Financial Stability Challenges".

[2] Different from fiat-backed stablecoins, algorithmic stablecoins use a different pegging mechanism without physically holding a pool of reserve assets.

In this paper, we analyze the economics of US dollar fiat-backed stablecoins and shed light on the possibility and probability of stablecoin runs. Stablecoins are uniquely designed with features of both exchange-traded funds (ETFs) and money market funds (MMFs). The majority of investors trade stablecoins in competitive secondary markets. Fluctuations in the secondary market price reflect changes in demand and supply but do not involve any direct fire sale of assets similar to fluctuations in the price of ETF shares. Asset sales only occur when the stablecoin issuer meets redemption requests in the primary market. The issuer liquidates some of its assets to pay $1 in cash for each stablecoin redeemed similar to MMFs, but redemption requests can only be submitted by a limited number of arbitrageurs approved by the stablecoin issuer. These arbitrageurs buy stablecoins trading below $1 in secondary markets to redeem them for $1 in primary markets, which allows them to pocket arbitrage profits while providing liquidity to investors.

Despite stablecoins' unique design and their tradability in competitive secondary markets, we show that they remain vulnerable to panic runs by investors in the spirit of Diamond and Dybvig (1983). This is because the fixed $1 redemption price in the primary market reinstates run incentives among secondary market investors, who fear that APs will retract from providing liquidity to them if the stablecoin issuer can no longer honor the $1 redemption value.

Interestingly, the concentration of arbitrageurs embeds an inherent tradeoff between run risk and price stability. If issuers only approve a small number of arbitrageurs to redeem tokens for cash in primary markets, arbitrage is less efficient and the same selling pressure would depress prices more in secondary markets. However, precisely because sellers in secondary markets would receive lower prices, their "first-mover advantage" from selling stablecoins in a run decreases. In other words, approving more arbitrageurs for more efficient arbitrage would exacerbate run risk and be counterproductive for financial stability.

More specifically, our dataset of fiat-backed stablecoins is constructed as follows. We collect transaction-level data on each stablecoin creation and redemption event for the 6 largest fiat-backed stablecoins: Tether (USDT), Circle USD Coin (USDC), Binance USD (BUSD), Paxos (USDP), TrueUSD (TUSD), and Gemini dollar (GUSD), on the Ethereum, Avalanche, and Tron blockchains. We obtain this data from each blockchain by converting transaction-level blockchain data into a usable format.

For each stablecoin, we also extract average trading prices in secondary markets from the main exchanges. Further, we obtain the composition of reserve assets for USDT and USDC, which reported these breakdowns at various points in 2021 and 2022.

From our novel data, we observe that the concentration of arbitrageurs in the primary market, where stablecoins are directly redeemed for cash with issuers, varies across stablecoins. For example, USDT only has 6 arbitrageurs redeeming shares during the average month and the largest arbitrageur accounts for 64% of the total redemption activity. In contrast, the arbitrage market at USDC is more competitive with 521 active arbitrageurs in an average month. We also find that trading prices in the secondary market for stablecoins frequently deviate from zero with discounts occurring 27.2% to 41.6% of the time and premia occurring 57.3% to 72.8% of the time. We note that these price deviations are not analogous to money market funds "breaking the buck" nor are they an indicator of runs. Rather, stablecoin prices fall below $1 when secondary market investors' selling pressure is not fully absorbed by the arbitrageurs, who purchase stablecoins in secondary markets and redeem them for $1 each in primary markets. In this sense, stablecoins trading below $1 is similar to ETF shares trading at a discount to their NAVs.[3]

We further observe that stablecoins with fewer arbitrageurs have higher average discounts in secondary markets. For example, the average discount at USDT is 55bps, while the average discount at USDC is only 1bps. At the same time, USDT also has more illiquid assets, like corporate bonds and loans, as part of their reserve assets than USDC. These observations leave open the question of how stablecoins choose the concentration of their arbitrageurs and how the choice relates to their asset illiquidity. After all, if approving more arbitrageurs can minimize secondary price deviations, then why wouldn't all stablecoin issuers simply allow for a competitive arbitrage market?

We develop a model to rationalize our empirical observations, assess the potential for stablecoin runs, and analyze the effect of market structure. Our theory applies a Diamond-Dybvig-style model to stablecoins and characterizes its unique design with features of both ETFs and MMFs. There are three types of agents: investors, arbitrageurs, and a stablecoin issuer. Specifically, investors are endowed

[3]The parallel to "breaking the buck" at money market funds would be a failure by stablecoin issuers to honor the $1 redemption value in primary markets, which has not yet materialized thus far.

with stablecoins that are aimed at providing a fixed value and backed by an illiquid reserve asset. They may sell stablecoins to arbitrageurs in the secondary market but they cannot directly redeem them from the issuer, similar to the case of ETFs. Arbitrageurs bid in a double auction to absorb any residual selling pressure from investors, and can then redeem stablecoins with the issuer in the primary market for one dollar, which resembles the redemption of MMFs shares. To honor the fixed redemption price of one dollar, the issuer liquidates its illiquid reserve asset pre-maturely until it defaults, after which only the liquidation value will be paid to redeeming arbitrageurs.

Our model shows that panic runs by investors on stablecoins can happen despite investors only being able to sell stablecoins in the secondary exchange at the market price. The conventional view is that exchange-traded claims like stocks and ETF shares are less runnable than bank deposits because the trading price falls as more investors sell, which creates a natural strategic substitutability. In the context of stablecoins, however, arbitrageurs are promised a fixed redemption price by the issuer. Hence, investors who choose to hold stablecoins may end up getting a less valuable stablecoin in the future because they bear the costs induced by the issuer's firesales of illiquid assets while meeting arbitrageurs' redemptions at \$1. In this way, stablecoins' fixed primary market price re-introduces strategic complementarity among secondary market investors.

We endogenize the run probability using global games to evaluate the effect of arbitrage efficiency. Surprisingly, we find that the run risk of stablecoins decreases in the concentration and increases in the balance sheet capacity of the arbitrage market. To understand this result, note that investors compare the benefit of selling their shares in the secondary market, i.e., the secondary market price, to the benefit of remaining invested in the stablecoin in the long run. When arbitrage is more efficient, price stability in the secondary market is higher because arbitrageurs are more willing to absorb selling pressure from investors. This higher trading price increases the benefit of selling stablecoins for investors and thereby amplifies their first-mover advantage when they expect other investors to sell. In contrast, when arbitrage is less efficient, small quantities of investor sales can have a substantial impact on stablecoin prices so the risk of secondary price discounts is higher. This price impact of stablecoin sales discourages panic selling and thereby mitigates run risk. Therefore, arbitrageurs act as a firewall between the

primary and secondary markets for stablecoins, which induces a trade-off between stablecoins' price stability and run risk.

Stablecoin issuers optimally design the structure of their arbitrage sectors to trade off price stability and run risk. Recall that USDT holds more illiquid assets than USDC while also approving fewer arbitrageurs than USDC. This is consistent with USDT restricting entry into their arbitrage market to partially offset the increased run-risk from holding more illiquid reserve assets.

Our model further provides an analytical solution for stablecoins' run probability, which we calibrate to quantify the run risk of the two largest stablecoins, Tether (USDT) and USD Coin (USDC). Our first input is the elasticity of redemptions in the primary market. Based on the model, redemption volumes should be more responsive to deviations in the secondary market price when there is a larger number of arbitrageurs. Empirically, we regress daily discounts against daily redemption volumes normalized by the total outstanding volume for each stablecoin. We find that the coefficient for USDT is larger in absolute magnitude than for USDC, which is consistent with the higher AP concentration of USDT constraining redemption volume to be less sensitive to price dislocations. Magnitude-wise, a 10 percentage point higher redemption volume corresponds to a 3.0 cent larger discount at USDT and a 1.3 cent larger discount at USDC.

To measure the overall illiquidity of USDT and USDC's reserve portfolios, we calculate the average discounts of their reserve assets weighted by their portfolio weights. We follow Bai, Krishnamurthy and Weymuller (2018) to proxy asset discounts with collateral haircuts by asset class. Intuitively, more liquid assets are more readily pledged to obtain cash at short notice while more illiquid assets incur a higher discount. On average, the reserve assets of USDT are more illiquid than those of USDC, but both of them shift towards holding more liquid assets over the sample period.

We estimate the distribution of the probability at which the risky asset payoff does not materialize. We use CDS spreads to evaluate the expected recovery value of each portfolio component and then calculate how the expected recovery value of the stablecoin issuer's overall reserve portfolio varies over time using historical data. The resulting empirical distribution is close to but not only concentrated at 1, consistent with USDT and USDC holding mostly but not exclusively safe assets.

Finally, we use our model to quantify the run risk at the two largest US dollar stablecoins. Tether and Circle, which make up the bulk of the market at $76.4 billion and $37.7 billion in January 2022. To calibrate our model parameters, we construct a novel dataset comprising of stablecoins' primary market transactions, secondary market trades, and reserve asset composition. Our estimates imply a higher run risk for USDT, the largest stablecoin, compared to USDC, the second-largest stablecoin, due to higher liquidity transformation. However, USDC also processes significant run risk due to less concentrated arbitrage and more concentrated deposit holdings.

Our paper contributes to a large literature on runs and liquidity transformation (e.g, Diamond and Dybvig, 1983, Allen and Gale, 1998, Bernardo and Welch, 2004, Goldstein and Pauzner, 2005). It has also been shown that MMFs are subject to panic runs because their shares are redeemed by investors at a fixed price (Kacperczyk and Schnabl, 2013, Parlatore, 2016, Schmidt, Timmermann and Wermers, 2016), while closed-end funds and ETFs are typically viewed as less runnable because their shares are tradable at market prices without direct liquidation of the underlying assets (Jacklin, 1987, Allen and Gale, 2004, Koont, Ma, Pastor and Zeng, 2021). By carefully modeling the unique combination of ETFs and MMFs in the design of stablecoins, we show that panic runs may still happen despite their trading on competitive secondary markets and investors' inability to access primary markets.

Methodologically, our estimation of run risks is enabled by the use of the global games approach to derive a unique run threshold. Goldstein and Pauzner (2005) shows that in the classic Diamond and Dybvig (1983) bank setting global strategic complementarity fails and thus the standard global games approach (e.g., Morris and Shin, 1998) does not directly apply, but a unique threshold run equilibrium exists as long as there is one-sided strategic complementary. In the stablecoin setting where the secondary and primary markets are separated, even one-sided strategic complementarity may not hold because selling the first unit of stablecoin in the secondary market generates a first-order price impact. However, we are able to show that a unique threshold run equilibrium still exists, which provides a foundation for our calibration to quantify stablecoin run risks. Relatedly, Egan, Hortacsu and Matvos (2017) build a structural model to quantify bank instability, highlighting the feedback between endogenous bank default and deposit withdrawals.

We also contribute to the emerging stablecoin literature by analyzing and quantifying the run risk of US dollar stablecoins. Several recent papers explore runs on algorithmic stablecoins (e.g., Adams and Ibert, 2022, Uhlig, 2022) after the Terra-Luna crash in 2022. Closely related to our work is Liu, Makarov and Schoar (2023), who examine the dynamics of the Terra USD run, focusing on how investor characteristics affect run behavior and the financial inclusion implications. On fiat-backed stablecoins, Barthelemy, Gardin and Nguyen (2021) and Liao and Caramichael (2022) analyze the potential impact of fiat-backed stablecoin activities on the real economy. Frost, Shin, Wierts (2020), Gorton and Zhang (2021), and Gorton, Ross and Ross (2022) compare stablecoins to the banking sector pre-deposit-insurance. Griffin and Shams (2020) suggest that, prior to 2020, Tether was used to manipulate Bitcoin prices. Lyons and Viswanath-Natraj (2021) show that USDT's creation and redemption respond to price deviations and who relate stablecoin price stability to defending exchange rate pegs. Kim (2022) finds that increases in the issuance of USDT and USDC lead to decreases in Treasury and commercial paper yields. Kozhan and Viswanath-Natraj (2021) analyze DAI, which is a stablecoin overcollateralized by risky non-USD assets. Li and Mayer (2021), d'Avernas, Maurin, and Vandeweyer (2022) and Routledge and Zetlin-Jones (2022) are theoretical papers on the mechanisms stablecoins use to maintain peg stability, encompassing algorithmic and collateral-backed stablecoins as well as fiat-backed stables. We provide a complementary yet distinct perspective of stablecoins as financial intermediaries engaged in liquidity transformation. Through this lens, we highlight the possibility of panic runs and relate run risk to the design of the primary market.

Our paper also fits more broadly into the literature on cryptocurrencies and decentralized finance, discussed and surveyed in Harvey, Ramachandran and Santoro (2021), John, Kogan and Saleh (2022), and Makarov and Schoar (2022).

The rest of the paper proceeds as follows. Section 2 describes institutional details of the stablecoin market and Section 3 explains the data we use. Section 4 documents several empirical facts that motivate our model in Section 5. Section 6 explains the model calibration and results. Finally, Section 7 concludes.

2 Institutional Details

Stablecoins are blockchain assets whose value is claimed to be stable at $1. Blockchain assets can be self-custodial: a user can use crypto wallet software, such as Metamask, to hold, send, and receive stablecoins directly. These tokens are not stored with any trusted intermediary: rather, a "private key" – a long numeric code, kept only on the user's hardware device – is used to prove to the blockchain network that the user owns her tokens and to direct the network to take actions such as transfer tokens to other wallets. Others have no access to individuals' private keys so they have no ability to take funds from individuals' wallets. Stablecoins are thus a useful way to hold US dollar assets in settings where there is a lack of trusted financial intermediaries that can be relied on to custody US dollar assets on behalf of market participants.

Relative to other blockchain assets like bitcoins, the defining feature of stablecoins is (relative) price stability. The largest stablecoin issuers attempt to achieve price stability by promising to back each stablecoin token by at least $1 in off-blockchain US dollar assets. These fiat-backed stablecoins have experienced a rapid expansion over the last few years. Within two years' time, the total asset size of the six largest fiat-biased stablecoins has grown from $5.6 billion at the beginning of 2020 to exceed $130 billion at the beginning of 2022 (Figure 2). The largest stablecoin is Tether (USDT), which made up more than 50% of the total market size at $76.4 billion in January 2022. Circle USD Coin (USDC) and Binance USD (BUSD) are second and third at $37.7 and $14.4 billion. Paxos, (PUSD), TrueUSD (TUSD), and Gemini dollar (GUSD) are significantly smaller with a market size of around or below $1 billion. The asset size of fiat-backed stablecoins has experienced ups and downs in 2022 but remains high at $136 billion in June 2022.

We proceed to explain the design of stablecoins and how they attempt to achieve price stability. A diagram illustrating the primary and secondary market for stablecoins is shown in Figure 1.

2.1 The Primary Market

Stablecoin tokens are created/minted and redeemed/destroyed in the primary market with US dollar cash as shown on the left-hand side of Figure 1. To create a stablecoin token, an arbitrageur sends $1 US dollar to the issuer, through a bank transfer or other means; the issuer then sends a stablecoin token into the market participant's crypto wallet. Analogously, to redeem a stablecoin token, for each stablecoin token that the market participant sends to the issuer's crypto wallet, the issuer sends $1 US dollar, for example through a bank transfer, into the market participant's bank account. The primary market for stablecoins resembles a money market fund in the traditional financial system. Please see Appendix A for further details.

Importantly, not all market participants can freely become arbitrageurs to participate in the redemption and creation of stablecoin tokens in the primary market. Stablecoin issuers differ in how easily and costly market participants can access primary markets. For example, while USDC allows general businesses to register as arbitrageurs and charges no fees for redemptions and creations, USDT restricts AP registration, imposes a minimum transaction size of $100,000, and charges the greater of 0.1% and $1000 per redemption. USDT also requires a lengthy due-diligence process and imposes restrictions on where arbitrageurs can be domiciled.

2.2 The Secondary Market

The majority of market participants trade existing stablecoins for fiat currencies in secondary markets, as shown on the right-hand side of Figure 1. Crypto-exchanges like Binance allows customers to make US dollar deposits, and then trade US dollars for USDT, USDC, or BUSD with other market participants.[4] The price of stablecoin tokens in the secondary market is thus driven by the demand from stablecoin buyers and the supply from stablecoin sellers. When there is a surge in stablecoin sellers on the secondary market, the secondary market price would drop but the closed-end nature implies that there are no direct liquidations of any reserve assets involved. In this way, the buying and selling of stablecoins on secondary markets resemble the trading of ETF shares on the exchange.

[4]Please see Appendix A for details regarding the use of different crypto exchanges.

2.3 Shock Transmission from the Secondary to the Primary Market

Nevertheless, selling pressure in the secondary market for stablecoins can spill over to affect the primary market through arbitrageurs. When investor selling pressure in the secondary market depresses stablecoin prices to be below $1, arbitrageurs can profit from purchasing stablecoin tokens for below $1 in secondary markets, and redeeming them one-for-one for $1 with the stablecoin issuer in primary markets as long as the issuer does not default. Analogously, if positive demand shocks in secondary markets caused stablecoins to trade above $1, arbitrageurs could profit from creating stablecoin tokens one-for-one for dollars in primary markets and then selling them at higher prices in secondary markets. Thus, the $1 redemption value of stablecoins in primary markets pulls the trading price of stablecoins towards $1 in secondary markets through the trading incentive of arbitrageurs.

At the same time, this arbitrage process implies that investor selling pressure in secondary markets can eventually trigger fire sales of assets when stablecoin issuers liquidate reserves to meet arbitrageurs' $1 redemption in cash. These fire sales can become especially costly if large amounts of redemptions occur in a short period of time and if illiquid reserve assets can only be converted to cash at a discount. If redemptions and discounts are large enough, the issuer may fail to pay the promised $1 for each stablecoin token redeemed, and the stablecoin defaults.

2.4 Uses of Stablecoins

Stablecoins have a number of uses. First, they are a fairly low-cost way to transact in US-dollar assets. As of January 2023, sending tokens on the Ethereum costs around $1 per transaction, and transactions finalize in under a minute.

Stablecoins are also being used as a store of value and medium of exchange in settings where inflation is high, capital controls and financial repression are prevalent, and trust in intermediaries is low. For example, humanitarian organizations have used stablecoins to circumvent banking fees and regulatory frictions.[5] Some firms in Africa have begun using stablecoins for international payments to

[5] See Fortune.com and Rest Of World.

suppliers in Asia.[6] In settings with high inflation, such as Lebanon and Argentina, individuals have begun storing value and transacting using stablecoins.[7] Some merchants in these areas have begun accepting stablecoins as a form of payment.[8]

Finally, stablecoins are used with other smart contracts within the space of "decentralized finance." For example, market participants can use stablecoin tokens to purchase other blockchain tokens, such as ETH, MKR, or UNI, using an automated market maker protocol such as Uniswap. Market participants can also lend stablecoin tokens on lending and borrowing protocols, such as Aave and Maker, allowing them to receive positive interest rates, and also to use these assets as collateral to borrow other assets. In a way, stablecoins provide a safe store of value and medium of exchange resemble for the blockchain ecosystem similar to the role of deposits and money market fund shares in the traditional financial system.

3 Data

We compile a novel and comprehensive dataset that sheds light on stablecoins' on-chain primary market activity, secondary market prices, and reserve assets.

3.1 Primary Market Data

The core dataset used in our analysis is data on each stablecoin creation and redemption event for the 6 largest fiat-backed stablecoins: Tether (USDT), Circle USD Coin (USDC), Binance USD (BUSD), Paxos (USDP), TrueUSD (TUSD), and Gemini dollar (GUSD), on the Ethereum, Avalanche, and Tron blockchains. We obtain this data from each blockchain based on "chain explorer" websites, which process transaction-level blockchain data into a usable format. We use Etherscan for Ethereum, Snowtrace for Avalanche, and Tronscan for Tron.

[6] See Rest Of World.

[7] See CNBC and Rest Of World for a discussion of the Lebanon case, and Coindesk and EconTalk for a discussion of Argentina.

[8] For example, the Unicorn Coffee House in Beirut, Lebanon accepts USDT (Tether) as a form of payment.

As described in Section 2, there are two ways that stablecoin tokens can be minted or redeemed. First, the stablecoin's "mint" or "burn" functions can be called directly to the primary market participant's wallet. To capture this category of actions, we query Etherscan for all cases in which the "mint" and "burn" functions are called for each stablecoin. Second, the stablecoin issuer can send or receive stablecoins from their "treasury" address. To capture this category, we identify the treasury address or addresses for each stablecoin, and then query Etherscan for every send or receive transaction involving the treasury address. Logistically, some issuers, such as Tether, tend to mint a large quantity of stablecoin tokens into "treasury" addresses they control, then issue tokens to market participants simply by transferring tokens out of their treasury wallet; whereas other issuers, such as TrueUSD, occasionally directly mint stablecoin tokens into the wallet addresses of market participants. On the other hand, most issuers handle redemptions by having market participants send tokens to a treasury wallet address. If the treasury wallet has a large balance of redeemed stablecoins, the issuer will occasionally "burn" quantities of the stablecoin, removing them from the technical outstanding balance of the token.[9]

Using our data extraction process, we see, for each stablecoin creation and redemption event, the precise timestamp of the event; the amount of the stablecoin redeemed or created; and the wallet address of the entity involved in stablecoin creation or redemptions. We also observe the "gas" fee – that is, the transaction fee paid to Ethereum miners for including the transaction in the blockchain – paid for each transaction. Some wallet addresses are tagged on Etherscan, as they are known to belong to large entities such as crypto exchanges. Using Etherscan wallet tags, we are able to group some wallets that are known to belong to the same economic entity.

We calculate the total issued market capitalization of a given stablecoin at any point in time, as the total technical market capitalization of the stablecoin, minus the amount of the stablecoin held in "treasury" addresses. This is because tokens held in treasury wallets need not be backed one-to-one by US dollars, and thus should not count as part of the total market capitalization of stablecoins in circulation.

[9]The exception to this rule is that TrueUSD occasionally handles redemptions by "burning" tokens directly from market participants' wallets, rather than the treasury.

3.2 Secondary Market Data

For each of the 6 stablecoins in our data, we extract the hourly closing prices for trades from the main exchanges, including Binance, Bitfinex, Bitstamp, Bittrex, Gemini, Kraken, Coinbase, Alterdice, Bequant, and Cexio. In our main analysis, we calculate daily prices for each stablecoin as the weighted average of hourly closing prices across these exchanges, where the weights are by trading volume. Differences in stablecoin prices across the main exchanges are generally negligible, hence the price series are not substantially affected by the weights we put on different exchanges.

3.3 Reserves

Stablecoins' reserve assets are not recorded on the blockchain. However, USDT and USDC reported breakdowns of their reserve assets at various points in 2021 and 2022 as part of their balance sheets. We obtain these breakdowns for USDT and USDC. The other four stablecoins have not released breakdowns of their reserve asset composition but state the broad categories of their reserves. We obtain and discuss these asset types in the next section.

4 Stylized Facts

In this section, we present a set of new stylized facts about stablecoins, which informs our model and calibration to quantify the run risk of stablecoins.

4.1 Secondary Market Price

Fact 1. *The trading price of stablecoins in the secondary market commonly deviates from $1. This price deviation per se does not constitute a run by investors.*

Figure 3 shows the price at which different stablecoins trade on the secondary market over time. We observe that the secondary market price rarely stays fixed at $1. Rather, stablecoins trade at a discount

to $1 27.2% to 41.6% of the time and trade at a premium to $1 57.3% to 72.8% of the time for our sample of stablecoins (see Table 2).

The extent of these price deviations varies by stablecoin. While the average discount at USDT is 55bps, the average discount at USDC is only 1bps. The average discount of BUSD, TUSD, and USDP are also below that of USDT at 1bps, 11bps, and 18bps, respectively, while that of GUSD is the highest at 78bps. The median discounts are generally smaller in magnitude than the average discounts, but the variation in the cross-section remains similar. The average and median premia also show significant variation in the cross-section.

The trading of stablecoins at a discount to $1 has been commonly associated with "breaking the buck" as in the case of money market funds and even as evidence for panic runs.[10] We note that these are misconceptions. Stablecoins maintaining a "stable value" of $1 refers to the amount that primary market participants receive or pay when they redeem existing stablecoins or create new stablecoins with the stablecoin issuer. The notion of "breaking the buck" thus corresponds to primary market participants not receiving a full $1. This scenario has not yet occurred at any of the stablecoins in our sample despite their secondary market price frequently deviating below $1. The secondary market price is the trading price of stablecoins on exchanges. It is essentially the share price of a closed-end fund and analogous to the share price of an ETF. Just like ETF prices can deviate from the NAV of the underlying portfolio, stablecoin prices can deviate from $1. This stablecoin price falling below $1 simply captures the selling pressure of stablecoins in the secondary market and is not a direct indicator of "breaking the buck" or panic runs.

4.2 Primary Market Concentration

Fact 2. *The redemption of stablecoins in the primary market is performed by a set of arbitrageurs, whose concentration varies by stablecoin.*

[10] For example, see https://www.nytimes.com/2022/06/17/technology/tether-stablecoin-cryptocurrency.html and https://www.cnbc.com/2022/05/17/tether-usdt-redemptions-fuel-fears-about-stablecoins-backing.html

Table 3 shows the characteristics of daily primary market redemption activity on the Ethereum blockchain for different stablecoins. We observe that on an average day, USDT only has 2 APs engaged in redemptions, whereas USDC has 33. The concentration of APs' market shares also varies. The largest AP at USDT performs 93% of all redemption activity, while the largest AP at USDC performs 54%. BUSD, USDP, and TUSD lie in between USDT and USDC in terms of the number of redeeming APs and AP concentration. GUSD has the most concentrated AP market with one AP essentially being in charge of all redemptions.

We repeat the analysis at the monthly level in Table 4. The monthly snapshot may better capture the market structure of the primary market than the daily snapshot if not all APs are active every day. Indeed, we observe that the number of APs redeeming stablecoins is larger at the monthly level. However, the AP market remains highly concentrated for USDT, with only 6 APs redeeming shares during the average month and the largest AP accounting for 64% of the total redemption activity. In contrast, USDC has 521 active APs in an average month but the top 1 and top 5 APs make up 45% and 85% of all redemption activity. As before, USDP, and TUSD lie in between USDT and USDC in terms of the number of redeeming APs and AP concentration. GUSD has the most concentrated AP market at the monthly level as well.

Further, notice that in the average month, the volume of redemptions at USDT is $615 million, while that at USDC is $2976 million. In comparison, the total volume of outstanding tokens at USDT was 1.5 to 2 times of that of USDC. Thus, the larger number and lower concentration of APs at USDC is correlated with a higher volume of redemptions relative to the total asset size as well.

In the appendix, we repeat Tables 3 and 4 for the Tron and Avalanche blockchains and obtain similar variations in AP concentration across stablecoins.

4.3 Secondary Market Price and Primary Market Concentration

Fact 3. *Stablecoins with a more concentrated set of arbitrageurs experience more pronounced discounts in the secondary market.*

We proceed to analyze the potential effects of AP concentration. We calculate the average monthly discount and the average number of redeeming APs for each stablecoin and plot them in Figure 4a. A clear negative trend emerges: stablecoins with fewer APs, like USDT and GUSD, have higher average discounts in their secondary market prices than stablecoins with more APs, like USDC and BUSD. Another way to capture AP concentration is through the market share of the largest APs. In Figure 4b, we repeat the analysis with the market share of the top 5 APs. The relationship is positive. Stablecoins whose top 5 APs consistently perform a larger share of total redemptions, like USDT and GUSD, have higher average discounts than other stablecoins with lower AP concentration. In other words, it seems that higher AP competition is associated with reduced price dislocations in secondary markets.

One question arising from this trend is why some stablecoins choose to have a more concentrated AP sector. If AP competition can indeed stabilize secondary market prices, all stablecoins should be incentivized to open up AP access and encourage the entry of new APs. In our model, we show that a counteracting force is the presence of panic runs by investors, which are more likely with a more competitive AP sector. We show that the probability of panic runs is especially pronounced if the reserve assets are more illiquid, which makes AP concentration even costlier. In the next subsection, we illustrate that USDT indeed also has more illiquid reserve assets.

4.4 Liquidity Transformation

Fact 4. *Stablecoins engage in varying degrees of liquidity transformation by investing in illiquid assets.*

Stablecoins are not literally backed by US dollars in the form of cash. Rather, they hold USD-denominated assets with varying degrees of illiquidity as reserves. Table 1 shows the composition of reserve assets for USDT and USDC on reporting dates. Overall, reserve assets of both USDT and USDC are far from being fully liquid, with those of USDT being more illiquid.

A significant portion of reserve assets is in the form of deposits and money market instruments. In June 2021, these two asset classes took up 60.7% and 59.5% of reserve assets at USDT and USDC, respectively. Money market instruments include commercial paper and certificates of deposits. For USDT, deposits include "cash deposits at financial institutions and call deposits, i.e., deposits that may

be withdrawn with two days' notice or less; fiduciary deposits, i.e., deposits made by banks on behalf of and for the benefit of members of the consolidated group; and, term deposits, i.e., deposits placed by members of the consolidated group at its banks for a fixed term." For USDC, deposits include "US dollar deposits at banks and short-term, highly liquid investments that are readily convertible to known amounts of cash and have a maturity of less than or equal to 90 days from purchase." Thus, except for deposits in checking accounts, money market instruments and other types of deposits are not fully liquid, i.e., they can not be freely converted to cash at short notice. For example, time deposits and certificates of deposit experience a discount when demanded before their maturity date.

USDT also holds a significant portion of reserves in the form of Treasury bills, which increased from 24.3% in June to 47.6% in March 2022. In contrast, USDC reduced its Treasury holdings from 15.0% in July 2021 to 0% in August 2021. USDC states that their Treasuries include "US government treasury bills, notes and bonds with a maximum maturity of 3 years". While Treasuries are generally liquid and safe security, the extent of their liquidity varies by type and over time. For example, on-the-run Treasuries and Treasury bills are much more liquid than off-the-run Treasuries and non-bills.

The remaining reserve assets are comprised of more illiquid assets, including municipal and agency securities, foreign securities, corporate bonds, corporate loans, and other securities. USDT still held a sizable amount of these illiquid assets in March 2022, with 4.5%, 3.8%, and 6.0% in corporate bonds, corporate loans, and other assets, respectively. While the exact identity of other assets is not disclosed, it is mentioned that they do include crypto investments. In June 2021, USDC held 0.4%, 15.9%, and 9.5% in municipal and agency securities, foreign securities, and corporate bonds respectively. USDC's holding of these assets is reported to have dropped to zero starting in September 2021, with all assets held in the form of the deposits described above.

The other four stablecoins do not publish reserve breakdowns, but they report that their assets are limited to deposits, Treasuries, and money market instruments. For example, a statement issued by BUSD and USDP in July 2021 claims that they hold 96% of cash equivalents and 4% of Treasury bills. GUSD states that their reserves are "held and maintained at State Street Bank and Trust Company, Signature Bank, and within a money market fund managed by Goldman Sachs Asset Management, invested only in U.S. Treasury Obligations." TUSD also claims that their US dollar balance is held by

"U.S. depository institutions and Hong Kong depository institutions" and that they "include US dollar cash and cash equivalents that include short-term, highly liquid investments of sufficient credit quality that are readily convertible to know amounts of cash."

5 Model

In this section, we build a model to analyze the potential for stablecoin runs. The model aims for achieving three goals. First, the model formulates the notion of runs on the primary market of stablecoins and explicitly derives the run probability, linking it to the level of stablecoin liquidity transformation and the concentration of arbitragers. Second, the model formulates the stablecoin issuer's optimal design of its primary market structure. Finally, the model allows us to quantify the run risks for a number of major stablecoins.

5.1 Setting

The economy has three dates, $t = 0, 1, 2$, with no time discount. There are three groups of risk-neutral players, 1) a competitive fringe of identical, infinitesimal investors indexed by i, 2) a sector of $n \geq 3$ arbitrageurs or APs, and 3) a stablecoin issuer.

At $t = 0$, investors are born; each investor would incur a cost of c_i, which follows a distribution function $G(c)$, to participate in the stablecoin market. Once participated, each investor is endowed with one stablecoin. An investor participates only when its participation cost is smaller than its expected utility from participation, which will be determined in equilibrium. There are two types of assets: the dollar, which is also the consumption good and serves as the numeraire, and an illiquid and potentially productive reserve asset. The stablecoin is initially backed by the reserve asset held by the issuer. The initial value of the reserve asset is normalized to one dollar. We introduce the features of the two assets shortly below. The stablecoin issuer may also choose n at $t = 0$, that is, design the structure of its primary market, to maximize its expected profit, which we introduce in Section 5.4. Before that, we

take n as exogenous and focus on investors' equilibrium behaviors after they have participated, and as such, we also normalize the population of participating investors to one.

Participating investors are subject to idiosyncratic liquidity shocks at $t = 1$ as in Diamond and Dybvig (1983). Each investor is uncertain about her preferences over consumption at $t = 1$ and $t = 2$. At the beginning of $t = 1$, an investor learns her preferences privately: with probability $\pi > 0$ she is an early-type and gets utility from date-1 consumption only, while with probability $1 - \pi$ she is a late-type and gets utility from consumption from both dates.

At $t = 1$, a total of $\lambda \geq \pi$ investors decide to sell stablecoins to APs in a secondary market at the market price p in exchange for a consumption good called dollar, where both λ and p will be determined in equilibrium. Dollar is riskless, liquid, and it serves as the numeraire. Households cannot directly redeem the stablecoin for dollar from the issuer, but APs are able to redeem the stablecoins from the issuer in a primary market, getting a fixed price of one dollar per stablecoin if the issuer is solvent. To raise dollars to meet AP redemptions at $t = 1$, the issuer has to liquidate the illiquid reserve asset prematurely at a liquidation cost of $\phi \in (0, 1]$, that is, liquidating one unit of asset yields $1 - \phi$ dollar only. Hence, the issuer is solvent if and only if $\lambda < 1 - \phi$. We assume $\pi < 1 - \phi$ to rule the uninteresting case that early investors alone render the issuer default. When $\lambda \geq 1 - \phi$, the issuer defaults, and the redeeming APs will get the liquidation value per total stablecoins redeemed, that is $(1 - \phi)/\lambda$. Expecting the amount of dollars to be redeemed from the issuer, APs bid in a double auction (e.g., in the manner of Kyle (1989) and Du and Zhu (2017)) to buy the stablecoins from λ selling investors. Denote the AP sector's total balance sheet capacity in the auction by S. The auction determines the secondary-market price p, the magnitude of which reflects the de-pegging risk of the stablecoin.

At $t = 2$, the economy entails aggregate risk. With probability $p(\theta)$, the economy enters a good state: the reserve asset matures and yields a value of $R(\phi) \geq 1$ dollar. With probability $1 - p(\theta)$, the economy enters a bad state: the reserve asset fails and yields zero. Here, $R(\phi)$ is decreasing in ϕ and $p(\theta)$ is increasing in $\theta \in \Theta$. We call θ the fundamentals of the economy which captures the level of aggregate risk, which is unknown to investors, APs, or the stablecoin issuer before $t = 2$. Intuitively, the reserve asset is more likely to profit as the fundamentals are better, and its long-term maturing value $R(\phi)$ increases in its illiquidity, capturing a notion of liquidity risk premium.

In the good state of the economy, participating investors and the stablecoin issuer share the value of the reserve asset based on the following rule. Unlike a security, a stablecoin never pays dividends. Thus, the net maturing gain of the reserve asset, that is, $R(\phi) - 1$, only gets accrued to the stablecoin issuer but not the investors. Rather, the remaining $1 - \lambda$ investors consume the initial value 1 per unit of the remaining reserve asset, plus a convenience value $\eta > 0$ per stablecoin at $t = 2$. Beyond the stablecoin, investors cannot access the underlying asset market or any other investment technology to transfer wealth across time.

To endogenize investors' stablecoin selling decisions and hence the stablecoin's run risk, we follow the global games literature to assume that each investor i obtains a private signal $\theta_i = \theta + \varepsilon_i$ at $t = 1$, where the noise term ε_i are independently and uniformly distributed over $[-\varepsilon, \varepsilon]$. As usual in the literature (e.g., as in Goldstein and Pauzner (2005)), we focus on arbitrarily small noise in the sense that $\varepsilon \to 0$, but the model results also hold beyond the limit case. An investor's selling decision depends on the signal that she obtains. Note that we do not impose any restrictions on the distributions of p, θ, or the increasing function $p(\theta)$, which would conveniently allow us to map the model to any empirical distribution of fundamentals.

5.2 Discussion of Model Specification

Before proceeding, it is useful to discuss several important modeling choices to highlight the economics underlying the model. The discussion also highlights in what sense our model parsimoniously captures the most important features of the stablecoin markets.

First, our model purposefully features the separate but connected primary and secondary markets of stablecoins, as discussed in Section 2 and observed in Section 4. This separation is important for us to separately define the de-pegging and run risks of stablecoins and to analyze the relationship between these two types of risks. In reality, most retail investors of stablecoins cannot directly participate in the stablecoin creation and redemption process with the issuers, which is captured by investors only being accessible to the secondary market. Given our focus on stablecoin selling and redemptions, any excess supply for stablecoins by investors in the secondary market then must be met by AP redemptions

with the issuer in the primary market. Our modeling of the AP's activity as a double auction with a fixed redemption price from the issuer closely mirrors the real-world redemption and destroy process of stablecoins, which is only available to the small set of APs.

It is also instructive to draw connections to ETFs and MMFs to further highlight the uniqueness of the stablecoin market and how our model captures this uniqueness. Like stablecoins, ETFs also feature the separation of the primary and secondary markets in that only APs can access the primary market and any excess demand or supply of ETF shares from investors in the secondary market must be met by APs (e.g., see Koont, Ma, Pastor and Zeng (2021) for a model of ETFs highlighting these features). However, ETFs notably differ from stablecoins in that AP creations and redemptions are predominantly performed in-kind with the issuer, that is, APs are delivered the underlying assets rather than cash when redeeming ETF shares. In contrast, a stablecoin AP gets a fixed amount of one dollar when redeeming one unit of stablecoin with the issuer provided the stablecoin issuer is solvent. This key difference thus resembles MMFs before the 2016 Money Market Reform in that MMF investors also get a fixed amount of one dollar in redemption provided the issuer is solvent (e.g., see Parlatore (2016) for a model of MMFs). Note that, however, MMF shares are not tradable in any secondary markets. Hence, stablecoins uniquely combine the two-layer market structure of ETFs and the fixed-value in-cash redemption feature of MMFs, and our model parsimoniously captures this combination.

Our model captures liquidity transformation and concentration of APs as the two most important economic sources of variation across different stablecoins, as documented in Section 4. Liquidity transformation is captured by the illiquidity cost parameter ϕ. Below, we call ϕ as either liquidation cost or haircut. Economically, ϕ captures that the liquidation of loans and bonds in secondary markets can depress their prices (see Duffie, 2010, for a review). It may also capture negative real impacts when liquidations of loans and bonds affect the capacity of governments and corporates rolling over their debt (e.g. He and Xiong, 2012). The concentration of APs is captured by the parameter n, holding the AP sector's total balance sheet capacity in bidding fixed. This specification helps us separately consider the effects of AP concentration and AP balance sheet costs on de-pegging and run risks.

The model also parsimoniously captures how the stablecoin issuer and coin-holding investors share the long-term value of the stablecoin. As of now, stablecoins do not pay any dividends to investors and

thus are not regulated as securities by the U.S. SEC. Hence, the net return of the underlying reserve assets is only accrued to the issuer. However, by holding stablecoins in the long term (rather than selling them to meet immediate liquidity demand), participating investors enjoy a convenience yield that is currently specific to the use of stablecoins as a payment method in crypto investments and decentralized finance contexts, and potentially beyond as a widely adopted means of payment going forward. As we further specify in Section 5.4, the expected revenue of the stablecoin issuer hinges on this specific form of value sharing between the issuer and investors.

Finally, our model follows the global games approach to endogenize the run risk of stablecoins, which is the focus of this paper. One key assumption of the global games approach is the information structure: the fundamentals are unobservable but each agent obtains a private signal about them. We view this assumption to be plausible for the stablecoin market because of its opacity: essentially no stablecoin issuers disclose asset-level information about their reserves. On the other hand, investors in the stablecoin market are likely to be more sophisticated than those in more traditional financial markets, justifying their ability to obtain private and heterogeneous signals about the fundamentals.

5.3 Equilibrium Analysis

We first solve for the equilibrium secondary-market stablecoin price at $t = 1$, p, when λ investors choose to sell. Define

$$K = \frac{1}{S}\frac{n-1}{n-2}\,, \tag{5.1}$$

and impose the following parametric assumption to ensure that the secondary-market price is positive:

$$1 - \phi - K > 0\,. \tag{5.2}$$

We have the following result:

Proposition 1. *The stablecoin's secondary-market price at $t = 1$ is given by*

$$q(\lambda) = \begin{cases} 1 - K\lambda & \lambda \leq 1 - \phi\,, \\ \dfrac{1-\phi}{\lambda} - K\lambda & \lambda > 1 - \phi\,. \end{cases} \tag{5.3}$$

Proposition 1 shows that the stablecoin's secondary-market price depends on selling pressure λ, and further, the level of liquidity transformation ϕ, AP balance sheet capacity S, and AP concentration n. Specifically, q is decreasing in λ and ϕ while increasing in S and n. All these comparative statics are intuitive. A higher selling pressure λ depresses stablecoin price due to the standard excess supply effect, leading to higher de-pegging. A higher level of liquidity transformation ϕ does not affect the stablecoin price when the issuer is solvent but translates to a lower stablecoin price when the issuer defaults because of the lower liquidation value. Indeed, a higher ϕ also makes the issuer more likely to default. A higher AP balance sheet capacity implies that APs are more willing to bid to absorb the selling pressure, supporting a higher secondary-market price. Finally, a less concentrated AP sector, that is, a higher n implies that APs bid more competitively, also leading to a higher secondary-market price. Looking forward, we will show that these features play an important role in determining the relationship between secondary-market de-pegging risk and primary-market run risk of stablecoins.

Viewing the stablecoin's secondary-market price q as a function of λ specifically, we highlight two important features of $q(\lambda)$. First, it is strictly decreasing in λ everywhere. Second, it features a kink at $\lambda = 1 - \phi$, that is, when the stablecoin issuer just defaults. The first feature points to the standard notion of strategic substitutability usually present in many financial markets including the ETF market: the more investors sell, the lower the price is, making an investor less likely to sell. However, we show that the second feature, that is, the jump in price due to the issuer's inability to keep the fixed redemption price as it defaults, which resembles MMFs, may eventually give rise to a strong enough first-mover advantage in selling, as we analyze later.

Now we consider the late investors' decision of selling the stablecoin at $t = 1$ or not (recall that early investors always sell). In making this decision, a late investor compares the secondary-market price $q(\lambda)$ she may get by selling the stablecoin at $t = 1$ to the return she may get at $t = 2$ if she does

not sell, which is given by

$$v(\lambda) = \begin{cases} p(\theta)\left(\dfrac{1-\phi-\lambda}{(1-\phi)(1-\lambda)} + \eta\right) & \lambda \leq 1-\phi\,, \\ 0 & \lambda > 1-\phi\,. \end{cases} \tag{5.4}$$

To see why it is this case, notice that the issuer needs to liquidate

$$l(\lambda) = \begin{cases} \dfrac{\lambda}{1-\phi} & \lambda \leq 1-\phi\,, \\ 1 & \lambda > 1-\phi\,. \end{cases}$$

unit of the reserve asset to meet AP redemptions at $t = 1$, and only $1 - l(\lambda)$ unit remains at $t = 2$, whose value will be shared by the remaining $1 - \lambda$ late investors.

It is useful to compare the date-2 stablecoin value (5.4) to the date-1 secondary-market stablecoin price (5.3) and define a late investor's payoff gain of waiting until $t = 2$ versus selling at $t = 1$:

$$h(\lambda) = v(\lambda) - q(\lambda) = \begin{cases} p(\theta)\left(\dfrac{1-\phi-\lambda}{(1-\phi)(1-\lambda)} + \eta\right) - 1 + K\lambda & \lambda \leq 1-\phi\,, \\ -\dfrac{1-\phi}{\lambda} + K\lambda & \lambda > 1-\phi\,. \end{cases} \tag{5.5}$$

It is easy to see that $h(0) \geq 0$ when $p(\theta)$ is sufficiently large while $h(1) < 0$, implying that the model has multiple equilibria when θ is sufficiently large and if θ is common knowledge.

The intuition behind a stablecoin run can be further illustrated by Figure 5, which plots the function $h(\lambda)$. It is clear from Figure 5 that $h(\lambda)$ first increases in λ, then decreases, and then increases in λ again. The first region where $h(\lambda)$ increases reflects strategic substitutability arising from the secondary market of stablecoins. Because selling investors generate a price impact on the secondary market and depress the secondary-market stablecoin price, a late investor may find it less appealing to sell if other late investors sell if the price impact is sufficiently large. This force works against the standard first-mover advantage that is typically present in a bank run model like Diamond and Dybvig (1983) and acts to prevent a run from happening. However, because APs redeem stablecoins from the issuer at

a fixed price of \$1, the cost that waiting investors have to bear will increase as more and more late investors choose to sell. This force will ultimately dominate, offsetting the secondary-market strategic substitutability and leading to a decreasing $h(\lambda)$ in the second region. Eventually, the second force dominates as λ becomes sufficiently large, pushing $h(\lambda)$ to be negative, which reinstalls the first-mover advantage and leads to runs in equilibrium.

Under the global games framework, we have the following result:

Proposition 2. *There exists a unique threshold equilibrium in which late investors sell the stablecoins if they obtain a signal below threshold θ^* and do not sell otherwise.*

Proposition 2 implies that the model with investors' private and noisy signals has a unique threshold equilibrium. A late investor's selling decision is uniquely determined by her signal: she sells the stablecoin at $t = 1$ if and only if her signal is below a certain threshold. Given the existence of the unique run threshold, we can show that it satisfies the following Laplace equation:

$$\int_{\pi}^{1-\phi} (1 - K\lambda)\, d\lambda + \int_{1-\phi}^{1} \left(\frac{1-\phi}{\lambda} - K\lambda \right) d\lambda = \int_{\pi}^{1-\phi} p(\theta^*) \left(\frac{1-\phi-\lambda}{(1-\phi)(1-\lambda)} + \eta \right) d\lambda \,. \tag{5.6}$$

Solving the Laplace equation gives the following result:

Proposition 3. *The run threshold is given by*

$$p(\theta^*) = \frac{(1-\phi)(2-2\phi-2\pi-2(1-\phi)\ln(1-\phi)-(1-\pi^2)K)}{2\left((1+\eta(1-\phi))(1-\phi-\pi)+\phi\ln\phi-\phi\ln(1-\pi)\right)} \,. \tag{5.7}$$

which satisfies the following properties:

i). The run threshold, that is, run risk, is increasing in ϕ if and only if $g(\phi) > K$,[11] *where $g(\phi)$ is continuous and strictly decreasing in ϕ over $(0, 1-\pi)$, and satisfies $\lim_{\phi\to 0} g(\phi) > 0$.*

[11] The explicit form of $g(\phi)$ is given by

$$g(\phi) = \frac{2\left((\pi+\phi-1)((\phi-1)(\eta(\phi-1)+1)+\pi-\ln\phi)+(\phi-1)\ln(1-\phi)(\pi(\eta(\phi-1)-2)-2\phi+(\phi+1)\ln\phi+2)+\ln(1-\pi)\left(\pi-\left(\phi^2-1\right)\ln(1-\phi)+\phi-1\right)\right)}{\left(1-\pi^2\right)\left((1-\phi)(1-\eta(1-\phi))-\pi+\ln\phi-\ln(1-\pi)\right)} \,.$$

ii). The run threshold, that is, run risk, is decreasing in K (that is, increasing in n and increasing in S).

Proposition 3 gives an analytical solution of the run threshold and presents intuitive comparative statics about the stablecoin's run risk with respect to the level of liquidity transformation and the organization of the AP sector.

Part i) of Proposition 3 shows that a higher level of stablecoin liquidity transformation leads to a higher run risk when $g(\phi) > K$. This condition may be satisfied when ϕ is not too large for a given K. Intuitively, when the stablecoin engages in a higher level of liquidity transformation in the sense that it holds less liquid reserve asset, the first-mover advantage among investors becomes larger because an investor who chooses not to sell would have to involuntarily bear a higher cost of liquidation induced by selling investors. This leads to a higher run risk. However, when the reserve asset is too illiquid, the first-mover advantage could be dampened because too few investors can enjoy it. This intuition can be understood from equation (5.5): investors enjoy the first-mover advantage only when $\lambda \leq 1 - \phi$, that is, $h(\lambda)$ takes the value in the first line of (5.5); too high a ϕ shrinks the region in which the first-mover advantage can be realized. Thus, further increasing the level of liquidity transformation when $g(\phi) < K$ will reduce the run risk. Looking forward, we confirm empirically in Section 6 that $g(\phi) > K$ indeed holds for the major stablecoins, suggesting that further increasing liquidity transformation will likely increase their run risks.

Part ii) of Proposition 3 shows that a more efficient AP sector in terms of less AP concentration and higher AP balance sheet capacity leads to higher run risk. To understand this more surprising result, note that the connection between stablecoins' secondary and primary markets implies a trade-off between de-pegging and run risks. A more efficient AP sector implies a lower de-pegging risk because the APs are more willing to absorb selling pressure from investors and thus more able to support a stable secondary market trading price. However, this means that APs will support a higher trading price to selling investors and subsequently redeem more stablecoins in the primary market, leading to a larger first-mover advantage and higher run risk. In contrast, to reduce run risk, the stablecoin issuer has to bear a higher de-pegging risk with a less stable secondary market price. In other words, when the AP sector is more efficient, the de-pegging risk is lower, because APs are more willing to absorb

selling pressure from investors and thus more able to support a stable secondary market trading price. However, run risk actually increases, because APs support a higher trading price to selling investors, increasing the first-mover advantage for stablecoin sellers. When the AP sector is less efficient, the de-pegging risk is higher, since small quantities of stablecoin sales can have a substantial impact on stablecoin prices. However, the price impact of stablecoin sales in fact decreases first-mover advantage and discourages "panic selling", contributing to decreasing run risk. In this sense, the AP sector acts as a firewall between stablecoins' secondary and primary markets, and the stablecoin issuer optimally designs the structure of its AP sector to trade off between these two risks.

The analytical solution given in Proposition 3 allows us to calibrate the model and quantify the run risks of the stablecoins in reality. To this end, we can easily translate the run threshold into an ex-ante run probability, with the additional input of the fundamental distribution $F(\theta)$. The following definition gives us a formal notion of run risk, which we use in the calibration exercise in Section 6.

Definition 1. *The ex-ante run probability of a stablecoin is given by*

$$\rho = \int_{p(\theta)<p(\theta^*)} dF(\theta)\,, \tag{5.8}$$

where $p(\theta^*)$ is given by (5.7) and $F(\theta)$ is the prior distribution of the fundamentals.

5.4 Optimal Design of the Stablecoin Primary Market

To further illustrate the idea of APs act as a firewall between stablecoins' secondary and primary markets and the trade-off between de-pegging and run risks, we further study the optimal design of the stablecoin primary market. Specifically, we focus on the optimal concentration of APs.

Given the potential for a panic run, the stablecoin issuer's design decision at $t = 0$ involves one key choice variable: n, that is, how many APs are allowed to perform primary-market redemptions. As described in Section 2, stablecoin issuers indeed consider the number of APs as one of the most important market design choices. We suppose that the stablecoin issuer chooses n to maximize its expected revenues at $t = 0$, which in turn depends on how many investors participate at $t = 0$. The

issuer's objective function is given by

$$\max_{n} E[\Pi] = \underbrace{G(E[W])}_{\substack{\text{population of} \\ \text{participating investors}}} \underbrace{\int_{p(\theta)\geq p(\theta^*)} \left(p(\theta)(R(\phi)-1)\frac{1-\phi-\pi}{1-\phi} \right) dF(\theta)}_{\substack{\text{expected issuer revenue per} \\ \text{participating investor}}}, \tag{5.9}$$

where each investor's expected utility of participation is given by

$$E[W] = \int_{p(\theta)<p(\theta^*)} q(1)dF(\theta) + \int_{p(\theta)\geq p(\theta^*)} (\pi q(\pi) + (1-\pi)v(\pi))\, dF(\theta)\,, \tag{5.10}$$

in which $q(\cdot)$ and $v(\cdot)$ are given by (5.3) and (5.4), which is in turn a function of θ, and $p(\theta^*)$ is given by (5.7) in Proposition 3.

The stablecoin issuer's objective function (5.9) intuitively captures its revenue base: it enjoys the net long-term return of the remaining reserve asset if no panic run happens (after possible liquidation to meet redemptions driven by early investors), and more participating investors allow the issuer to start with investing in more reserve assets. Turning to participating investors' expected utility $E[W]$, the first term in (5.10) denotes the expected welfare of all investors when a panic run happens, while the second term corresponds to the expected investor welfare when a run does not happen.

Solving the stablecoin issuer's problem (5.9), we have the following result about the stablecoin issuer's optimal choice of AP concentration:

Proposition 4. *When the stablecoin engages in a higher level of liquidity transformation, the stablecoin issuer optimally designs a more concentrated AP sector, that is, n^* decreases in ϕ when ϕ is not too large.*

Proposition 4 stems fundamentally from the trade-off between the de-pegging and run risks of stablecoins. Intuitively, when investors are subject to idiosyncratic liquidity risks, they do not know ex-ante whether they have to consume early or late. To attract more investor participation, the stablecoin allows investors to share their idiosyncratic risks by them jointly holding a pool of reserve assets and offering the ability to sell the stablecoin in the secondary market at a price potentially higher than what an investor would have gotten by holding the reserve assets herself. However, because of the run

risks, risk sharing may not always be achieved because everyone would just get the autarky outcome in a run scenario, hurting investors' expected utility and thus their participation as captured by the first term in (5.9). Further, a higher run risk also directly hurt the stablecoin issuer's expected revenue per participating investor as captured by the second term in (5.9), because the issuer would only enjoy the net long-term return of the reserve asset when no run happens. Thus, the issuer optimally accepts some level of de-pegging risk, that is, some deviation of the secondary-market price from its peg to avoid runs. This limits the ability of the stablecoin to provide immediate liquidity to early investors but would avoid a run. To achieve so, the issuer optimally chooses a concentrated AP sector to reduce the first-mover advantage among investors.

6 Model Calibration and Results

In this section, we calibrate our model to estimate run probability as defined in Definition 1. We start with a simple benchmark case of $\pi = 0$, which implies that the idiosyncratic shock of buying and selling stablecoins is mean zero and thus does not directly drive runs. This benchmark allows us to focus on panic runs and relate the run risk to the key stablecoin design features that we highlight. We focus our analysis on the largest two fiat-backed stablecoins, USDT and USDC, because of the availability of their reserve asset breakdowns.

We first explain our estimation of redemption elasticity, K, asset liquidity, ϕ, and the distribution of $p(\theta)$, before reporting the estimation results.

6.1 Elasticity of Redemptions in the Primary Market K

To estimate how responsive the volume of redemptions is to price discounts, we regress daily discounts against daily redemption volume for each stablecoin:

$$Discount_t = \beta Redemptions_t + FE_y, \tag{6.1}$$

where $Discount_t$ is the lowest observed secondary market price minus 1 on day t and $Redemptions_t$ is the volume of redemptions divided by the total outstanding volume of tokens on day t. We use the lowest secondary market price to better capture the extent of price dislocations that demand AP arbitrage rather than the price dislocations resulting from AP arbitrage. We normalize the volume of redemptions by the total outstanding volume of tokens to consider the difference in market sizes across stablecoins. Finally, we include a year fixed effect to capture potential structural shifts in the AP sector for each stablecoin. For example, the number and constraints of APs may evolve after some time with the growth of stablecoins.

Table 5 shows the results. We observe that the regression coefficients are negative for both USDT and USDC, which is consistent with larger redemption volumes on days with steeper discounts, i.e., more negative secondary market prices. Further, the coefficient for USDT is larger in absolute magnitude than for USDC, which is consistent with the higher AP concentration of USDT constraining redemption volume to be less sensitive to price dislocations. That is, a larger price dislocation is required to induce the same amount of redemptions for USDT than for USDC. Magnitude-wise, a 10 percentage point higher redemption volume as a fraction of the total volume outstanding corresponds to a 3.0 cent larger discount at USDT and a 1.3 cent larger discount at USDC.

6.2 Asset Illiquidity ϕ

We proxy asset illiquidity with haircuts following Bai, Krishnamurthy and Weymuller (2018) and Ma, Xiao and Zeng (2021). These haircuts proxy for the discount incurred when illiquid assets are converted into cash at short notice.[12] In other words, one minus the haircut is the amount of cash that stablecoin issuers can provide to APs redeeming at short notice by borrowing against the asset. More liquid assets are more readily pledged to obtain cash while more illiquid assets incur a higher discount. Figure 6 shows the median of asset discounts over time. In comparison, Treasuries are generally the most liquid, while corporate loans are the most illiquid.

[12]The New York Fed publishes haircuts on different securities when pledged as collateral in repo loans.

To measure the overall illiquidity of USDT and USDC's reserve portfolios, we calculate the average discounts of their reserve assets weighted by their portfolio weights. The results are shown in Figures 7a, and 7b. One challenge is that we do not know the liquidity of their deposits. As discussed in Section 4, these deposits include time deposits and CDs for which an early withdrawal penalty is incurred. These penalties generally range from half-years to two years' worth of interest rates, depending on the financial institution and contract length. We set the discount on the early withdrawal of deposits to be 0.5%. This is a relatively conservative measure given that the lowest asset discounts are at 2%. Further, 0.5% would have been the approximate one-year penalty rate on 5-year CDs in the latter half of 2021, which is the period for which asset breakdowns are available.

Overall, the reserve assets of USDT are more illiquid than those of USDC, but both of them shift towards holding more liquid assets over the sample period. The discount on USDT reserve assets decreased from 4.3% in September 2021 to 4.0% in March 2022. In comparison, the discount on USDC reserve assets drops from 0.9% in August 2021 to 0.5 % in September 2021. We use these estimates for the asset illiquidity parameter, ϕ, in our model.

6.3 Distribution of $p(\theta)$

Finally, our model requires us to take a stance on the distribution of $p(\theta)$, which is the signal of how likely the risky asset held in the issuer's portfolio is to pay nothing. To estimate p empirically, we use historical CDS prices to evaluate the extent to which the value of each portfolio component varies over time, allowing us to calculate a synthetic measure for how much the expected recovery value of the reserve portfolio is likely to fluctuate over time.

The CDS spread s_c on an asset class $c \in \{1 \dots C\}$ can be thought of as the probability of default under a recovery rate of 0. Since we assume 0 recovery rates in our model, for a single asset, s_c maps exactly to p in our model. Now, suppose the issuer holds a fraction q_c of her portfolio in asset class c. If each asset pays off 1 with probability s_c and 0 with probability $(1 - s_c)$, the portfolio as a whole has expected recovery value:

$$\sum_{c=1}^{C} (1 - s_c)\, q_c$$

We add an adjustment factor to account for the fact that stablecoin issuers tend to be overcollateralized. If the issuer holds $1+\xi$ in assets times the total number of stablecoin issued, then the expected recovery value of assets, for each unit of stablecoin issued, is:

$$p = (1+\xi) \sum_{c=1}^{C} (1 - s_c)\, q_c \tag{6.2}$$

Since p in the model is equal to the expected recovery value of assets per unit stablecoin issued, we will use (6.2) on each date we observe CDS spreads as one realization of p. We can think of (6.2) as the price of a composite security, which averages across CDS spreads of different components of a stablecoin issuer's portfolio, and accounts for the fact that issuers are slightly overcollateralized. With any set of CDS spreads on a given day, we can calculate a value of p using (6.2). By plugging CDS spreads from different dates into (6.2), we can calculate a distribution of signals p. Note that, when we plug CDS spreads into (6.2), we use spreads from a single day; hence, this method accounts for correlations between CDS prices of different asset classes.

We implement (6.2) using historical CDS spread data from Markit, from 2008 to 2022.

For deposits, we assign the average CDS of unsecured debt at the top 6 US banks to capture the riskiness of the banking sector.[13] We note that despite stablecoin issuers' claim that deposits are riskless in FDIC-insured institutions, they are not riskless or fully insured because deposit accounts exceeding 250K are not covered by deposit insurance. For Treasuries, we assign the CDS spreads on 3-year US treasuries. For money market instruments, we use CDX spreads on 1-year investment-grade corporate debt. For USDC's corporate bonds, we assign the 10-year investment-grade corporate CDX because they are stated to be of at least a BBB+ rating. For USDT's corporate bonds, we assign the average 10-year corporate CDX. The remaining categories, "foreign" and "other", do not have a clear mapping to the existing CDS series. For USDT, for example, assets in the "other" category include cryptocurrency, which could potentially be very risky. In our baseline results, we use the emerging market CDX spread as a proxy. We use the 10-year high-yield CDX spread as a robustness check.

[13] These include Bank of America, Wells Fargo, JP Morgan Chase, Citigroup, Goldman Sachs, and Morgan Stanley.

Table 6 shows the distributions of p for USDT and USDC on dates with reported balance sheets. The distributions of p are fairly concentrated near 1, with a narrow range from roughly 97% to 99.5%. In comparison, the distribution of p's for USDC is slightly worse than USDT, which arises from USDT's large holdings of Treasuries that have lower CDS spreads than bank deposits, which are the bulk of USDC's portfolio.

6.4 Calibration Results

Combining our estimates of the redemption elasticity, K and the asset illiquidity, ϕ, calculate run cutoffs according to (5.7) in Proposition 3. Then, we can infer run probabilities for each stablecoin in each time period based on the corresponding empirical distribution of the signal $p(\theta)$ following (5.8) in Definition 1.

The results for run probabilities are shown in Table 7. Overall, run probabilities are substantial. USDT's run probability was 3.45% in March 2022 and USDC's run probability was 0.14% in October 2021.

7 Conclusion

In this paper, we analyzed the possibility of panic runs on stablecoins. At a high level, stablecoins holders engage in liquidity transformation, offering APs the option to redeem stablecoins for cash dollars, while holding partially illiquid portfolios of assets. This creates the possibility for runs, where market participants sell tokens in secondary markets, leading APs to buy and redeem stablecoins for dollars with the issuer. We show, however, that stablecoin run risk is mediated by the market structure of the AP sector, which serves as a "firewall" between the secondary and primary markets. When the AP sector is more efficient, shocks in the secondary market transmit more effectively to the primary market; peg stability of stablecoins is thus improved, but the first-mover advantage for sellers is also higher, increasing run risk. If the AP sector is less efficient, shocks in secondary markets transmit less effectively; peg stability suffers, but run risk actually decreases, as the price impact of stablecoin trades

in secondary markets discourages market participants from panic selling. Calibrating the model to data, we quantified run risk for the two leading fiat-backed stablecoins by market cap.

References

ADAMS, AUSTIN and MARKUS IBERT (2022). "Runs on Algorithmic Stablecoins: Evidence from Iron, Titan, and Steel." Working Paper.

ALLEN, FRANKLIN and DOUGLAS GALE (1998). "Optimal Financial Crises." *Journal of Finance*, 53: 1245-1284.

ALLEN, FRANKLIN and DOUGLAS GALE (2004). "Financial Intermediaries and Markets." *Econometrica*, 72.4: 1023-1061.

AUER, RAPHAEL, JON FROST, LEONARDO GAMBACORTA, CYRIL MONNET, TARA RICE, and HYUN SONG SHIN (2022). "Central Bank Digital Currencies: Motives, Economic Implications, and the Research Frontier." *Annual Review of Economics*, 14: 697-721.

BAI, JENNIE, ARVIND KRISHNAMURTHY and CHARLES-HENRI WEYMULLER (2018). "Measuring Liquidity Mismatch in the Banking Sector." *Journal of Finance*, 73: 51-93.

BARTHELEMY, JEAN, PAUL GARDIN and BENOÎT NGUYEN (2018). "Stablecoins and the real economy." Working Paper.

BERNARDO, ANTONIO and IVO WELCH (2004). "Liquidity and Financial Market Runs." *Quarterly Journal of Economics*, 119: 135-158.

Brunnermeier, Markus, Harold James, and Jean-Pierre Landau. 2019. The Digitalization of Money. Working Paper.

COOPER, RUSSELL and THOMAS ROSS (1998). "Bank Runs: Liquidity Costs and Investment Distortions." *Journal of Financial Economics*, 41: 27-38.

D'AVERNAS, ADRIEN, VINCENT MAURIN, and QUENTIN VANDEWEYER (2022). "Can Stablecoins be Stable?" Working Paper.

DIAMOND, DOUGLAS and PHILIP DYBVIG (1983). "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy*, 91: 401-419.

DU, SONGZI and HAOXIANG ZHU (2017). "What is the Optimal Trading Frequency in Financial Markets?" *Review of Economic Studies*, 84: 1606-1651.

DUFFIE, DARRELL (2010). "Presidential Address: Asset Price Dynamics with Slow-Moving Capital." *Journal of Finance*, 65: 1237-1267.

DUFFIE, DARRELL (2019). "Digital Currencies and Fast Payment Systems: Disruption is Coming." Working Paper.

EGAN, MARK, ALI HORTACSU and GREGOR MATVOS (2017). "Deposit Competition and Financial Fragility: Evidence from the US Banking Sector." *American Economic Review*, 107: 169-216.

FROST, JON, HYUN SONG SHIN, and PETER WIERTS 2020. An early stablecoin? The Bank of Amsterdam and the governance of money. Working Paper.

GORTON, GARY, CHASE ROSS and SHARON ROSS (2022). "Making Money." Working Paper.

GORTON, GARY and JEFFERY ZHANG (2021). "Taming Wildcat Stablecoins." Working Paper.

GOLDSTEIN, ITAY and ADY PAUZNER (2005). "Demand Deposit Contracts and the Probability of Bank Runs." *Journal of Finance*, 60: 1293-1328.

GRIFFIN, JOHN and AMIN SHAMS (2020). "Is Bitcoin really untethered?" *Journal of Finance*, 75: 1913-1964.

HARVEY, CAMPBELL, ASHWIN RAMACHANDRAN, and JOEY SANTORO (2021). *DeFi and the Future of Finance*, 2021, John Wiley & Sons.

HE, ZHIGUO and WEI XIONG (2012). "Rollover Risk and Credit Risk." *Journal of Finance*, 67: 391-430.

JACKLIN, CHARLES (1987). "Demand Deposits, Trading Restrictions and Risk Sharing." in: Prescott, E. D., Wallace, N. (Eds.), *Contractual Arrangements for Intertemporal Trade*. Minnesota: University of Minnesota Press.

JOHN, KOSE, LEONID KOGAN and FAHAD SALEH (2022). "Smart Contracts and Decentralized Finance." Working Paper.

KACPERCZYK, MARCIN and PHILIPP SCHNABL (2013). "How Safe Are Money Market Funds?" *Quarterly Journal of Economics*, 128: 1073-1122.

KOONT, NAZ, YIMING MA, LUBOS PASTOR and YAO ZENG (2022). "Steering a Ship in Illiquid Waters: Active Management of Passive Funds." Working Paper.

KYLE, ALBERT (1989). "Informed Speculation with Imperfect Competition." *Review of Economic Studies*, 56: 317-355.

KOZHAN, ROMAN and GANESH VISWANATH-NATRAJ (2021). "Decentralized Stablecoins and Collateral Risk." Working Paper.

LI, YE and SIMON MAYER (2021). "Money creation in decentralized finance: A dynamic model of stablecoin and crypto shadow banking." Working Paper.

LIU, JIAGENG, IGOR MAKAROV and ANTOINETTE SCHOAR (2023). "Anatomy of a Run: The Terra Luna Crash." Working Paper.

LIAO, GORDON and JOHN CARAMICHAEL (2021). "Stablecoins: Growth Potential and Impact on Banking." Working Paper.

LYONS, RICHARD and GANESH VISWANATH-NATRAJ (2021). "What Keeps Stablecoins Stable?" Working Paper.

KIM, SANG RAE (2022). "How the Cryptocurrency Market is Connected to the Financial Market." Working Paper.

MA, YIMING, KAIRONG XIAO and YAO ZENG (2021). "Mutual Fund Liquidity Transformation and Reverse Flight to Liquidity." *Review of Financial Studies*, forthcoming.

MAKAROV, IGOR and ANTOINETTE SCHOAR (2022). "Cryptocurrencies and Decentralized Finance (DeFi)." *Brookings Papers on Economic Activity*, Spring 2022: 1-71.

MORRIS, STEPHEN and HYUN SHIN (1998). "Unique Equilibrium in a Model of Self-fulfilling Currency Attacks." *American Economic Review*, 88: 587-597.

MORRIS, STEPHEN and HYUN SHIN (2003). "Global Games: Theory and Applications." in Mathias Dewatripont, Lars Peter Hansen and Stephen J. Turnovsky (eds), *Advances in Economics and Econometrics (Proceedings of the Eighth Congress of the Econometric Society)*, 3: 56-114.

PARLATORE, CECILIA (2016). "Fragility in Money Market Funds: Sponsor Support and Regulation." *Journal of Financial Economics*, 121: 595-623.

ROUTLEDGE, BRYAN and ARIEL ZETLIN-JONES (2022). "Currency Stability Ssing Blockchain Technology." *Journal of Economic Dynamics and Control*, 142: 104-155.

SCHMIDT, LAWRENCE, ALLAN TIMMERMANN and RUSS WERMERS (2016). "Runs on Money Market Mutual Funds." *American Economic Review*, 106: 2625-2657.

UHLIG, HARALD (2022). "A Lunatic Stablecoin Crash." Working Paper.

Figure 1: The Design of Fiat-backed Stablecoins

This figure illustrates the design of fiat-backed stablecoins.

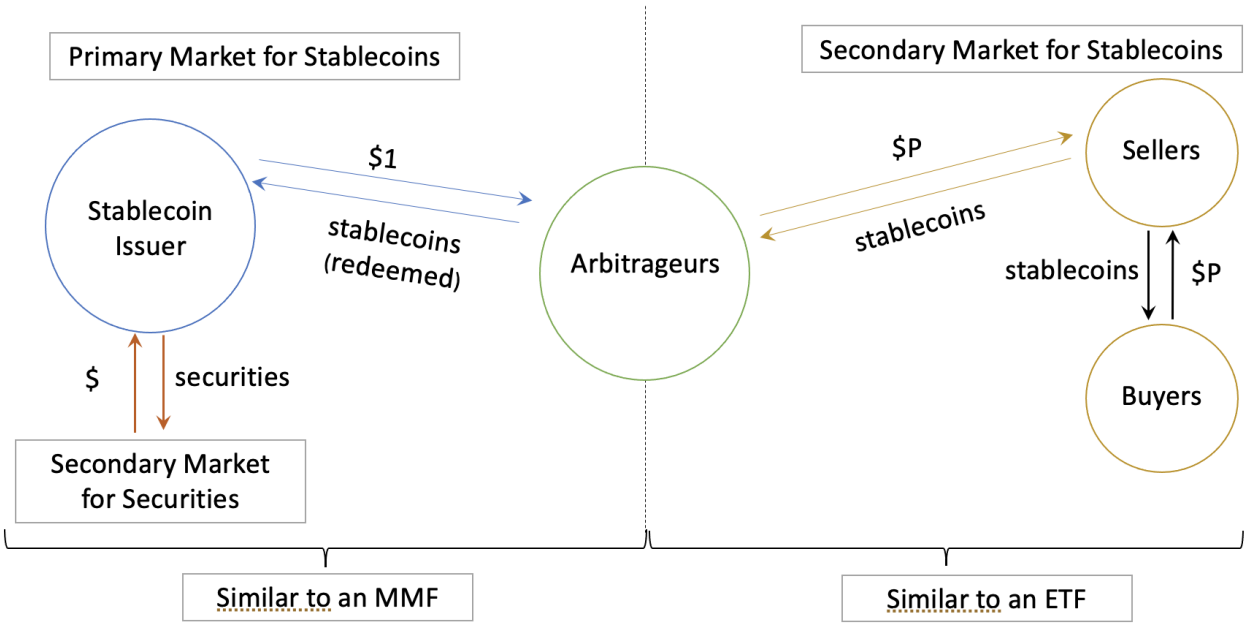


Figure 2: Asset Size of Fiat-backed Stablecoins

This figure shows the asset size of the six largest fiat-backed stablecoins over time.

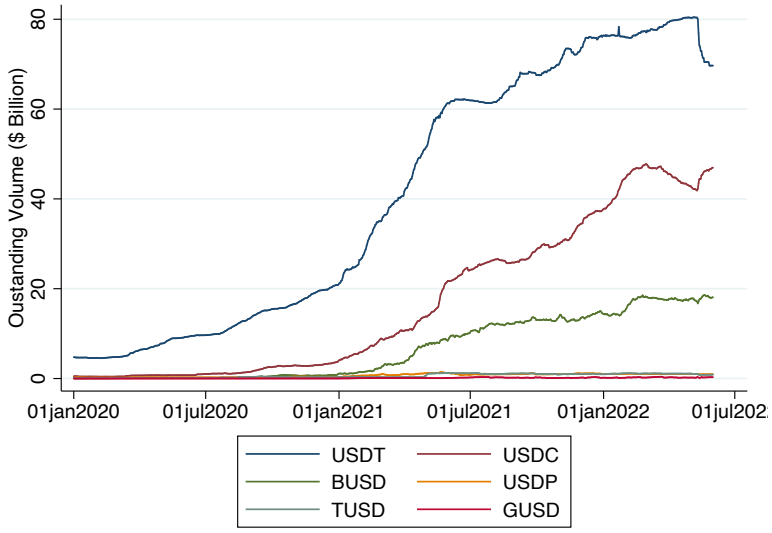


Figure 3: Secondary Market Trading Price

Panels (a) to (f) show the the daily secondary market trading price of USDT, USDC, BUSD, USDP, TUSD, and GUSD, respectively. Secondary market prices are volume-weighted average of trading prices from the exchanges listed in Section 2.

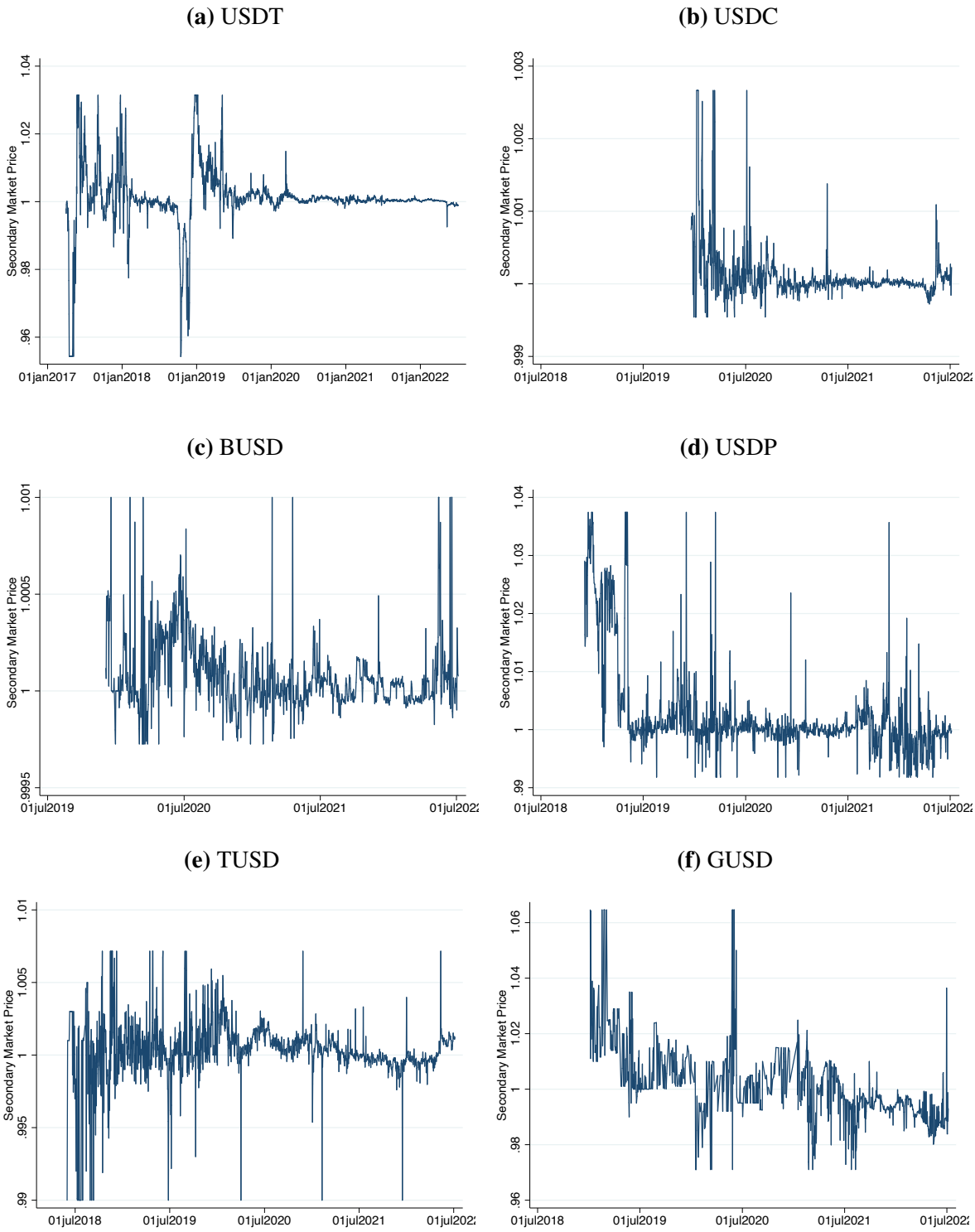
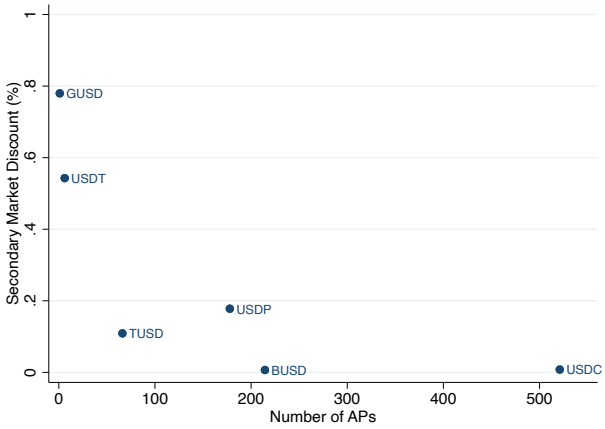


Figure 4: Secondary Market Discount and Primary Market Structure

This figure shows the relationship between secondary market price dislocations and primary market structure. In panel (a), each dot indicates the average secondary market discount and the average number of redeeming APs in a month for a given stablecoin. In panel (b), each dot indicates the average secondary market discount and the market share of the top give redeeming APs in a month for a given stablecoin.

(a) Number of APs



(b) Market Share of Top 5 APs

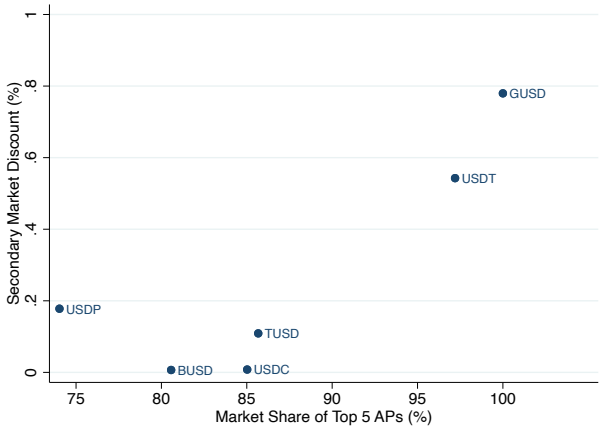


Figure 5: Payoff Gain of Late Investors

This figure shows a late investor's payoff gain between waiting until $t = 2$ versus selling at $t = 1$. Parameters: $p(\theta) = 0.97, \eta = 0.2, \phi = 0.05, K = 0.3$.

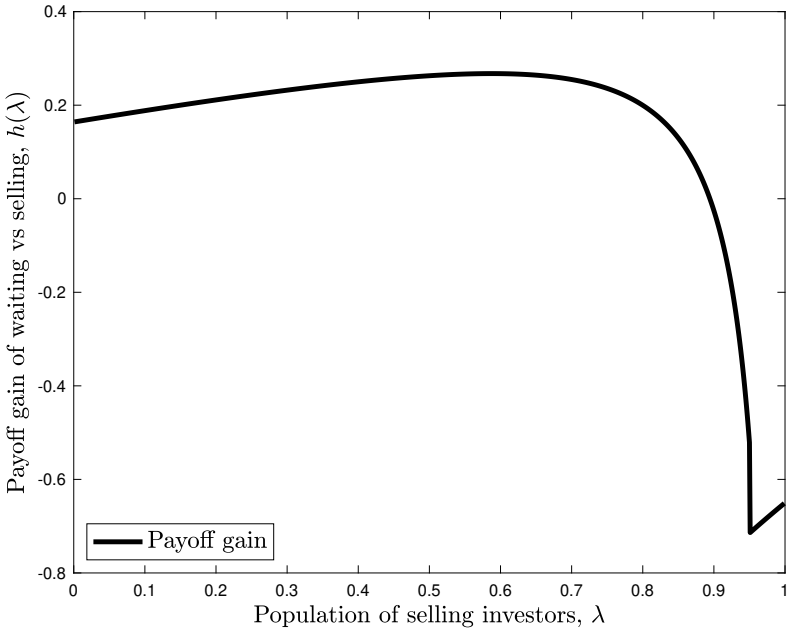


Figure 6: Liquidation Discounts

This figure shows median haircuts by collateral type. Data is from the New York Fed

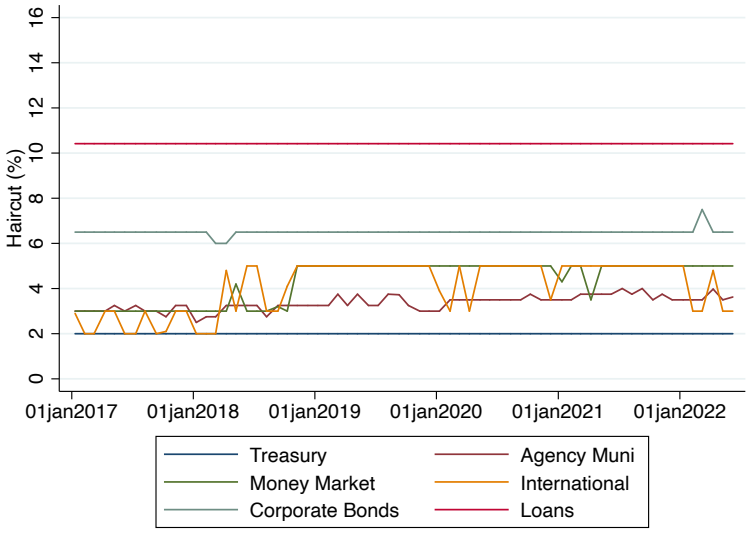
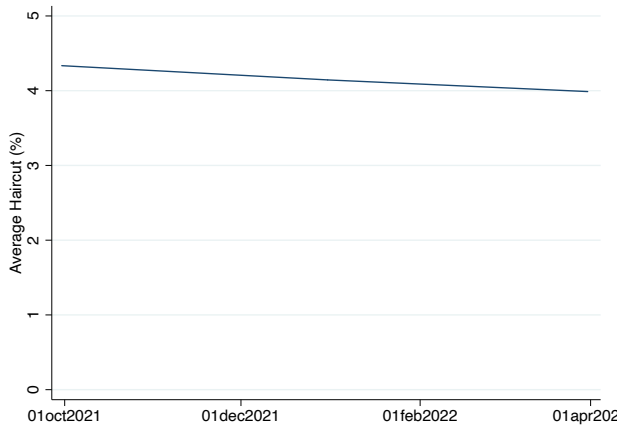


Figure 7: Asset Illiquidity

Panels (a) and (b) show the liquidation discount for USDT's and USDC's reserves. The sample period covers the dates for which a breakdown of reserve holdings for USDT and USDC overlapped.

(a) USDT Discount



(b) USDC Discount

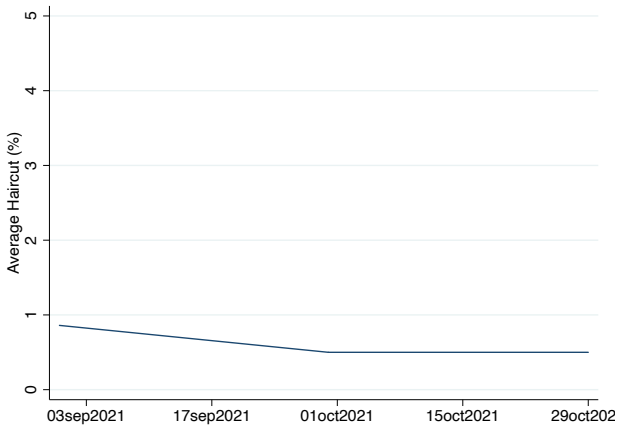


Table 1: Asset Composition

This table shows the breakdown of reserves by asset class for USDT and USDC. Data are available for the dates on which reserve breakdowns are published by USDT and USDC. For USDT, the "Deposit" category includes bank deposits, while for USDC, the "Deposit" category includes US dollar deposits at banks and short-term, highly liquid investments.

(a) USDT

	Deposits	Treas	Muni	MM	Corp	Loans	Others
2021/06	10.0	24.3	0.0	50.7	7.7	4.0	3.3
2021/09	10.5	28.1	0.0	45.7	5.2	5.0	5.5
2021/12	5.3	43.9	0.0	34.5	4.6	5.3	6.4
2022/03	5.0	47.6	0.0	32.8	4.5	3.8	6.4

(b) USDC

	Deposits	Treas	Muni	MM	Corp	Loans	Others
2021/05	60.4	12.2	0.5	22.1	5.0	0.0	0.0
2021/06	46.4	13.1	0.4	24.2	15.9	0.0	0.0
2021/07	47.4	12.4	0.7	23.0	16.4	0.0	0.0
2021/08	92.0	0.0	0.0	6.5	1.5	0.0	0.0
2021/09	100.0	0.0	0.0	0.0	0.0	0.0	0.0
2021/10	100.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2: Secondary Market Price and Volume

This table provides statistics about secondary market trading, including the average daily trading volume, the proportion of days with discounts and premiums, the average discount and premium, and the median discount and premium.

	USDT	USDC	BUSD	TUSD	USDP	GUSD
Average Daily Volume	16.4	15.4	13.5	11.4	10.5	7.3
Proportion of Discount Days (%)	30.5	27.2	34.9	38.2	41.6	39.7
Proportion of Premium Days (%)	69.5	72.8	64.4	61.4	57.3	58.9
Average Discount (%)	0.54	0.01	0.01	0.11	0.18	0.78
Average Premium (%)	0.36	0.02	0.02	0.13	0.64	1.17
Median Discount (%)	0.11	0.00	0.00	0.05	0.09	0.63
Median Premium (%)	0.11	0.01	0.01	0.10	0.18	0.82

Table 3: Primary Market Daily Redemption Activity

Panels (a) to (f) provide statistics about daily primary market redemption activity on the ethereum blockchain, including the number of APs, the market share of the top 1 and top 5 APs, and the volume of redemptions. For each variable, we show the average, 25^{th} percentile, 50^{th} percentile , and 75^{th} percentile of values across days in our sample.

(a) USDT

	mean	p25	p50	p75
AP Num	1	1	1	2
Top 1 Share	94	100	100	100
Top 5 Share	100	100	100	100
Vol (mil)	57	2	12	60

(b) USDC

	mean	p25	p50	p75
AP Num	33	8	14	28
Top 1 Share	54	45	50	59
Top 5 Share	96	95	98	100
Vol (mil)	103	2	15	134

(c) BUSD

	mean	p25	p50	p75
AP Num	21	8	15	28
Top 1 Share	59	40	56	76
Top 5 Share	94	90	96	100
Vol (mil)	62	8	27	82

(d) USDP

	mean	p25	p50	p75
AP Num	18	8	17	27
Top 1 Share	55	37	52	73
Top 5 Share	90	85	95	100
Vol (mil)	12	3	6	13

(e) TUSD

	mean	p25	p50	p75
AP Num	6	3	6	8
Top 1 Share	72	54	73	91
Top 5 Share	99	99	100	100
Vol (mil)	6	1	2	5

(f) GUSD

	mean	p25	p50	p75
AP Num	1	1	1	1
Top 1 Share	100	100	100	100
Top 5 Share	100	100	100	100
Vol (mil)	6	0	1	3

Table 4: Primary Market Monthly Redemption Activity

Panels (a) to (f) provide statistics about monthly primary market redemption activity on the ethereum blockchain, including the number of APs, the market share of the top 1 and top 5 APs, and the volume of redemptions. For each variable, we show the average, 25^{th} percentile, 50^{th} percentile , and 75^{th} percentile of values across months in our sample.

(a) USDT

	mean	p25	p50	p75
AP Num	6	3	6	8
Top 1 Share	66	42	61	89
Top 5 Share	97	98	100	100
Vol (mil)	577	46	123	763

(b) USDC

	mean	p25	p50	p75
AP Num	521	114	168	262
Top 1 Share	45	38	49	50
Top 5 Share	85	81	85	90
Vol (mil)	2976	160	460	4965

(c) BUSD

	mean	p25	p50	p75
AP Num	214	157	202	274
Top 1 Share	48	30	50	62
Top 5 Share	81	74	82	87
Vol (mil)	1596	233	1498	2720

(d) USDP

	mean	p25	p50	p75
AP Num	178	71	174	284
Top 1 Share	41	24	37	54
Top 5 Share	74	62	77	88
Vol (mil)	260	94	174	262

(e) TUSD

	mean	p25	p50	p75
AP Num	66	49	74	85
Top 1 Share	50	36	46	64
Top 5 Share	86	79	91	94
Vol (mil)	154	31	85	260

(f) GUSD

	mean	p25	p50	p75
AP Num	1	1	1	1
Top 1 Share	100	100	100	100
Top 5 Share	100	100	100	100
Vol (mil)	113	7	17	164

Table 5: Secondary Price Deviation versus Redemptions

This table shows the results from regressing the lowest daily secondary market price against the daily volume of redemptions for each USDT and USDC. The lowest secondary market price is the lowest hourly price for each coin on each day. The daily volume of redemptions is expressed as a proportion of the total outstanding volume of each stablecoin. We include a year fixed effect to account for structural shifts over time.

	USDT	USDC
	(1)	(2)
Redemption	-0.30**	-0.13**
	(0.13)	(0.06)
Observations	438	892
Adjusted R2	0.14	0.02

Table 6: Distribution of $p(\theta)$

This table shows quantiles of the distributions of the expected recovery value of assets per unit stablecoin. We combine Markit data on CDS spreads for different asset classes from 2008 to 2022, with data on stablecoin issuers' asset class holdings and over-collateralization ratios, using expression (6.2).

coin	date	p10	p25	p50	p75	p90
USDT	2021m9	0.9857	0.9896	0.9929	0.9940	0.9950
USDT	2021m12	0.9873	0.9908	0.9931	0.9941	0.9952
USDT	2022m3	0.9884	0.9915	0.9936	0.9945	0.9956
USDC	2021m8	0.9765	0.9858	0.9907	0.9931	0.9940
USDC	2021m9	0.9769	0.9861	0.9919	0.9944	0.9950
USDC	2021m10	0.9769	0.9861	0.9919	0.9944	0.9950

Table 7: Estimated Run Probabilities

This table shows our estimated run probabilities for different stablecoin issuers at different dates, calculated by combining our estimates of the distribution of $p(\theta)$, expected recovery value of assets per unit stablecoin using CDS data from expression (6.2), with the run cutoffs computed using expression (5.7).

month	coin	runprob
2022m3	USDT	0.0345
2021m10	USDC	0.0014

Table 8: Primary Market Daily Redemption Activity (Tron)

Panels (a) to (f) provide statistics about daily primary market redemption activity on the tron blockchain, including the number of APs, the market share of the top 1 and top 5 APs, and the volume of redemptions. For each variable, we show the average, 25^{th} percentile, 50^{th} percentile , and 75^{th} percentile of values across months in our sample.

(a) USDT

	mean	p25	p50	p75
AP Num	1	1	1	2
Top 1 Share	96	100	100	100
Top 5 Share	100	100	100	100
Vol (mil)	450	40	110	460

(b) USDC

	mean	p25	p50	p75
AP Num	33	7	17	28
Top 1 Share	67	45	67	94
Top 5 Share	93	91	98	100
Vol (mil)	2	0	0	2

(c) TUSD

	mean	p25	p50	p75
AP Num	1	1	1	1
Top 1 Share	97	100	100	100
Top 5 Share	100	100	100	100
Vol (mil)	10	0	0	2

Table 9: Primary Market Monthly Redemption Activity (Tron)

Panels (a) to (f) provide statistics about monthly primary market redemption activity on the tron blockchain, including the number of APs, the market share of the top 1 and top 5 APs, and the volume of redemptions. For each variable, we show the average, 25^{th} percentile, 50^{th} percentile , and 75^{th} percentile of values across months in our sample.

(a) USDT

	mean	p25	p50	p75
AP Num	5	2	4	6
Top 1 Share	72	53	68	94
Top 5 Share	100	100	100	100
Vol (mil)	4625	651	3575	7515

(b) USDC

	mean	p25	p50	p75
AP Num	446	11	317	391
Top 1 Share	58	33	51	81
Top 5 Share	84	78	85	100
Vol (mil)	41	3	24	70

(c) TUSD

	mean	p25	p50	p75
AP Num	4	2	3	7
Top 1 Share	87	69	95	100
Top 5 Share	100	100	100	100
Vol (mil)	61	0	21	32

Table 10: Primary Market Daily Redemption Activity (Avalanche)

Panels (a) to (f) provide statistics about daily primary market redemption activity on the avalanche blockchain, including the number of APs, the market share of the top 1 and top 5 APs, and the volume of redemptions. For each variable, we show the average, 25^{th} percentile, 50^{th} percentile , and 75^{th} percentile of values across months in our sample.

(a) USDT

	mean	p25	p50	p75
AP Num	1	1	1	1
Top 1 Share	100	100	100	100
Top 5 Share	100	100	100	100
Vol (mil)	31	5	30	60

(b) USDC

	mean	p25	p50	p75
AP Num	3	1	2	4
Top 1 Share	88	78	99	100
Top 5 Share	100	100	100	100
Vol (mil)	6	0	0	1

(c) BUSD

	mean	p25	p50	p75
AP Num	2	1	1	2
Top 1 Share	90	86	100	100
Top 5 Share	100	100	100	100
Vol (mil)	0	0	0	0

(d) TUSD

	mean	p25	p50	p75
AP Num	6	3	6	8
Top 1 Share	72	54	73	91
Top 5 Share	99	99	100	100
Vol (mil)	6	1	2	5

Table 11: Primary Market Monthly Redemption Activity (Avalanche)

Panels (a) to (f) provide statistics about monthly primary market redemption activity on the avalanche blockchain, including the number of APs, the market share of the top 1 and top 5 APs, and the volume of redemptions. For each variable, we show the average, 25^{th} percentile, 50^{th} percentile , and 75^{th} percentile of values across months in our sample.

(a) USDT

	mean	p25	p50	p75
AP Num	1	1	1	1
Top 1 Share	100	100	100	100
Top 5 Share	100	100	100	100
Vol (mil)	50	1	10	120

(b) USDC

	mean	p25	p50	p75
AP Num	34	18	32	47
Top 1 Share	49	31	42	60
Top 5 Share	94	87	96	99
Vol (mil)	111	3	16	219

(c) BUSD

	mean	p25	p50	p75
AP Num	22	10	18	30
Top 1 Share	37	30	40	42
Top 5 Share	83	73	82	94
Vol (mil)	0	0	0	0

(d) TUSD

	mean	p25	p50	p75
AP Num	66	49	74	85
Top 1 Share	50	36	46	64
Top 5 Share	86	79	91	94
Vol (mil)	154	31	85	260

A Appendix: Additional Institutional Details

A.1 Minting of Stablecoins

Practically, stablecoins are ERC-20 tokens. The stablecoin "smart contract," that is, the blockchain code that governs the behavior of the stablecoin, gives the stablecoin issuer the arbitrary right to create, or "mint", new stablecoin tokens, into arbitrary wallet addresses. Stablecoin issuers adopt technically slightly different strategies to issue and redeem stablecoins in primary markets. Some, like USDC, directly "mint" new coins using the token smart contract into customers' wallets. Others, like Tether, occasionally mint large amounts of stablecoin tokens to "treasury" wallets under their own control, and then issue stablecoins in primary markets by sending tokens from the "treasury" address to customers' wallets, and allow redemptions when customers send tokens to the treasury address.[14]

A.2 Trading on Crypto Exchanges

There are a number of ways individuals can purchase stablecoins with local fiat currency. One method is to deposit fiat on a custodial centralized crypto exchange (CEX), such as Binance or Coinbase. Centralized exchanges, like stock brokerages, keep custody of fiat and crypto assets on behalf of users, and allow users to purchase or sell crypto assets using fiat currencies. After purchasing stablecoins on a CEX, the user can then "withdraw" the stablecoins, instructing the CEX to send her stablecoins to a wallet address of her choosing, to self-custody the purchased stablecoins. Another approach is to use peer-to-peer exchanges, such as Paxful. On these platforms, users list offers to buy or sell stablecoins or other crypto tokens for other forms of payment. Accepted forms of payment in the US include Zelle, Paypal, Western Union, ApplePay, and many others. The exchange platform plays an escrow, insurance, and mediation role in these transactions. When a user buys a stablecoin, she sends funds to the exchange's escrow account and the stablecoin seller sends stablecoins to an address of the buyer's choosing. Once the buyer confirms receipt of the stablecoins, the exchange sends funds from

[14] Treasury address tokens technically count towards the market cap of any given stablecoin, but they are not economically meaningful as part of market cap, since Tether does not have to hold US dollar assets against tokens it holds in its treasury. Thus, we will not count tokens held in treasury addresses as part of the stablecoin supply in circulation.

the escrow account to the seller's account. In this process, purchased stablecoins are sent directly to the user's self-custodial wallet.

Sequential Search for Corporate Bonds*

Mahyar Kargar† Benjamin Lester‡ Sébastien Plante§
Pierre-Olivier Weill¶

January 30, 2023

Abstract

In over-the-counter (OTC) financial markets, customers search for trades by making repeated inquiries to dealers. Yet, there is little direct empirical evidence of this sequential search process, since existing transaction data only provides information about the times customers complete their trade, but no information about the times they search for a trade. In this paper, we shed new light on customers' sequential search process by leveraging a complete record of inquiries—successful and not—made on the leading electronic trading platform for corporate bonds. We obtain estimates of time to trade and trading costs, conditional on observable trade characteristics and the number of previously unsuccessful inquiries. We find that after the first failed inquiry, it takes two to three days for a customer to purchase an investment-grade bond. When interpreted through the lens of a sequential search model, our estimates highlight the importance of both observed *and unobserved* heterogeneity across customers. Overall, these estimates can serve as useful inputs into quantitative applications of search models and guide future theoretical explorations of sources of search frictions in OTC markets.

Keywords: Over-the-counter markets, corporate bonds

JEL Classification: G11, G12, G21.

*We thank Briana Chang, Tatyana Deryugina, Julia Fonseca, Mark Garmaise, Tim Johnson, Kumar Venkataraman, and conference and seminar participants at AEA/CEANA 2023, NBER Big Data and Securities Markets Spring 2023, INSEAD, UC Irvine, UCLA Macro-Finance Lunch, UC Riverside, and UIUC for comments and suggestions. We thank MarketAxess for providing the main data set used in this paper. The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. All errors are our own responsibility.

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

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

§University of Wisconsin-Madison. Email: splante@wisc.edu

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

1 Introduction

Over-the-counter (OTC) markets play a key role in the U.S. financial system: they include most fixed income securities, asset-backed securities, repurchase agreements, and various types of derivatives, along with a significant fraction of equity trading volume. OTC markets are different from exchange-based markets because they are decentralized: participants must first find a willing counterparty and then agree on the terms of trade. Search models have emerged as one of the workhorse theoretical and quantitative frameworks to analyze these markets, following the seminal work of Duffie, Gârleanu, and Pedersen (2005) (see Weill, 2020 for a survey). However, the main assumption that investors search for counterparties is not grounded in direct empirical evidence about the sequential trade process. The reason is simple: existing data from OTC markets is comprised of transaction records, which contain information about the time and price at which a trade occurs, but not about the time investors spend searching for a counterparty.[1]

In this paper, we leverage a proprietary data set to offer a unique window into the sequential search process of investors in one of the most studied OTC financial markets—the market for U.S. corporate bonds. The data provides a complete record of all inquiries made by customers, and the corresponding replies from dealers, on the leading electronic trading platform for corporate bonds, MarketAxess (MKTX). Crucially, by observing both successful *and unsuccessful* inquiries, the data allows us to estimate how long it takes a customer to execute a trade and how this length of time depends on the properties of the order and the characteristics of the customer. Moreover, by studying the behavior of both the customer and dealers over the course of the sequential search process, our analysis also offers new insights into the sources of delays in the trading process.

We start by documenting that inquiries fail to result in trade quite often—about a third of the time—which is consistent with the findings of Hendershott and Madhavan (2015) from an earlier time period. We go beyond this earlier work by analyzing the behavior of customers who, shortly after a failed inquiry, return to the market to make new inquiries for the same quantity of the same

[1]This limitation stands in stark contrast with other applications of search models. For example, data on unemployment spells is informative about workers' sequential job-finding process, while observations of time-on-the market for houses are informative about the sequential process for selling a home.

bond. In fact, by combining the data from MKTX with additional data from the Trade Reporting Compliance Engine (TRACE), we can observe when customers make additional electronic inquiries on MKTX for the same trade, when they complete the trade on MKTX, when they complete the trade *outside* of the electronic platform (via the traditional voice channels), and when they abandon the trade altogether.

Studying the details of the search process at such a granular level leads to novel estimates of the trading frictions that exist in the corporate bond market. For example, we find that it takes two to three days for a customer to complete the purchase of an investment-grade bond after an initial inquiry fails. Hence, given that approximately 70% of requests are filled at the first inquiry, a lower bound for the unconditional time to trade is about one day.[2] In addition, our analysis reveals that time to trade varies systematically across different characteristics of the order: time to trade is shorter for sells (relative to buys), small trades (relative to larger trades), and investment-grade bonds (relative to high-yield bonds). We also find that time to trade differs significantly across customers, as more "connected" investors get a larger number of responses to their inquiries and trade more quickly.

We also document the characteristics of contact rates and dealers' replies over the course of the sequential search process. We find that customers appear to make inquiries on MKTX more frequently as the number of failed inquiries increases, but these inquiries get fewer replies, the best offer gets worse, and the probability of trading falls. These dynamics could be an indication that the terms of trade worsen over the course of the search process, or they could reflect selection based on unobservables. We find evidence of the latter. When interpreted through the lens of a sequential search model, our estimates suggest that customers are heterogeneous in the intensity with which they make inquiries and in the number of responses they are able to elicit from dealers.

We believe that our analysis generates three main contributions to the existing literature. Our first contribution is to organize the data in a way that reveals customers' sequential search process. We do so by observing that when a customer becomes active on MKTX, she often submits a cluster

[2]This estimate is a lower bound because we do not have information about how long the customer was searching prior to submitting their first inquiry on MKTX.

of inquiries for a particular bond within a short period of time, rather than submitting one large inquiry. Following the practice in the equity market, we call this cluster a "parent order." Some of the inquiries in a given parent order are for different quantities of the same bond, a form of order splitting.[3] However, other inquiries can be identified as repeated attempts to trade a specific quantity; we call these "child orders." Because order splitting may be viewed as evidence of asymmetric information rather than search frictions, we focus on the sequential trade process for child rather than parent orders.

Summary statistics about child orders immediately reveal several interesting insights. First, sequential search is, indeed, a prominent feature of the trading process: as noted above, the probability that an inquiry is unsuccessful is about 30 percent, and customers routinely submit repeat inquiries. Conditional on the first inquiry failing, the median number of inquiries in a child order is 2, the 75th percentile is 3, and the 99th percentile is 9. Second, trade is non-exclusive: a customer may eventually trade on the voice market instead of MKTX, which we can observe using the enhanced version of the TRACE data set. The third takeaway is that the probability a child order ends without trade is significant, either because the investor abandons the trade altogether or creates a new child order by submitting an inquiry for a different amount. Lastly, organizing trades into parent and child orders has a significant effect on estimates of trading probabilities and time to trade. For example, we estimate that approximately 80% of child orders are eventually filled, whereas the existing literature (e.g., Hendershott and Madhavan, 2015) arrives at an estimate of approximately 67% when individual inquiries for trade are treated independently.

After organizing the data into parent and child orders, our second contribution is properly measuring the time it takes to successfully trade child orders. Even with our granular observations of inquiries and trade, measuring the time to trade remains a nontrivial exercise because of two potential sources of bias. The first is survivor bias created by "competing risks." For example, the measured average time to trade on MKTX is biased downwards because it is based on trades that have occurred relatively quickly, before the arrival of other events such as trading with a dealer outside

[3] Czech and Pintér (2020) provide evidence of informed investors splitting their orders across multiple dealers in the UK corporate bond market.

the platform or deciding to abandon this particular child order. The second concern is selection bias: since different types of inquiries (and customers) trade with different probabilities, we must account for potential changes in the composition of inquiries—both observable and unobservable—over the course of the sequential search process. We attempt to correct for these biases via Maximum Likelihood estimation: we assume that successful inquiries on MKTX, unsuccessful inquiries on MKTX, voice trades, and exit (i.e., abandoning the child order and/or beginning a new one) occur at independent exponential times with intensities that depend on the characteristics of the child order.

Equipped with these estimated intensities, we calculate the expected time it takes for a child order to trade, either on MKTX or via voice trade, in the event that the customer does not exit the child order. We find that it takes between two to three days to complete a trade *after a first failed inquiry*. Trade is much faster, by about a day, for sales than purchases. Block trades (with size above $5 million) take about one day longer to trade than micro-size ones (with size below $100,000). Bonds with amounts outstanding below the median take half a day more to trade. Bond age, turnover, and credit rating all have statistically significant impacts on time to trade, but with economically small effects. Customer connectedness, measured by the average number of responses that a particular customer elicits on MKTX, has an economically significant impact on time to trade. Moreover, we observe a large increase in time to trade, by more than a day, during the pandemic-induced crisis of March 2020, when the corporate bond market suffered severe liquidity disruptions (Kargar, Lester, Lindsay, Liu, Weill, and Zúñiga, 2021). Finally, comparing time to trade on MKTX and voice, we observe that child orders trade much faster on MKTX for all size categories except block trades, suggesting customers use the platform for execution quality rather than price discovery.

These estimates are helpful for at least two reasons. First, they can be directly applied to quantitative analyses based on search-theoretic models, since the arrival rates we estimate are crucial, yet controversial inputs that are typically identified via indirect inference. Second, the correlations we find between our estimates of time to trade and other observable outcomes provide

a natural starting point for additional empirical and theoretical work. For example, the fact that it takes longer for a customer to buy a bond than it does to sell it suggests that it might take time for dealers to locate (or "source") a bond. Alternatively, it could indicate that sellers are more distressed, on average, than buyers.

Our third contribution derives from studying the dependence of various trading outcomes on the number of previously unsuccessful inquiries, which helps us unpack the evolution of customers' and dealers' behavior over the course of the sequential search process. After controlling for various observed characteristics of child orders, we find that, with each additional failed inquiry, the number of dealer responses declines, the terms of trade deteriorate, and the time to trade increases. This dependence on the number of failed inquiries could, in principle, occur for two different reasons. First, dealers could adjust their behavior when they recognize a customer has attempted multiple, unsuccessful inquiries, perhaps because of information leakages or what Zhu (2011) calls the "ringing-phone curse." Alternatively, this dependence could derive from the fact that child orders differ in characteristics unobserved by the econometrician. We find evidence in support of the latter, rather than the former. In particular, when we repeat our estimation with child-order fixed effects, which control for all characteristics, observed and unobserved, the dependence of outcome variables on the number of failed inquiries largely disappears.

But what accounts for these child order fixed effects? In principle, it could be an unobserved characteristic of the customer or of the market at the time of the child order. We propose a simple theory that incorporates unobserved heterogeneity into the standard sequential search model in the tradition of McCall (1970), and argue that the dependence of outcome variables on the number of failed inquiries is consistent with child orders differing in their inquiry intensity and in the average number of responses they can elicit from dealers. More specifically, some child orders have low inquiry intensity and many responses, while others have high intensity and low responses.

Related literature

Our work is most closely related to the few other papers that have used the proprietary data from MKTX to analyze the impact of electronic trading on corporate bond market conditions (e.g., Hendershott and Madhavan, 2015; O'Hara and Zhou, 2021; Hendershott, Livdan, and Schürhoff, 2021). Our analysis differs from these papers in both our focus and our approach. More specifically, we are the first to organize the MKTX data into parent and child orders in order to offer new evidence about the sequential search process, including novel estimates of the time to trade conditional on the characteristics of the trade (size, direction, bond rating, and customer connectedness).

Our work also contributes to the vast empirical literature that studies corporate bond market liquidity based on transaction data. Some prominent examples include Schultz (2001), Bessembinder, Maxwell, and Venkataraman (2006), Edwards, Harris, and Piwowar (2007), Goldstein, Hotchkiss, and Sirri (2007), Bao, Pan, and Wang (2011), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018), and many others.[4] Our contribution relative to this literature is our attempt to measure the time to trade and to empirically investigate the sequential search process of customers in the corporate bond market. Our estimates of time to trade provide new empirical evidence on a dimension of liquidity at the center of search-theoretic models. Hendershott, Li, Livdan, and Schürhoff (2020) pursue similar goals but for a different dimension of liquidity (the cost of trade failures) in a different market (the market for collateralized loan obligations).

Our attempt to measure time to trade is related to earlier works in the OTC search literature which have proposed strategies to identify investors' search intensities. For example, according to the model of Afonso and Lagos (2015), Üslü (2019), and Brancaccio and Kang (2021), when search is random and the distribution over agents' state is continuous, every meeting results in a trade. This allows one to identify the search intensity from the trading intensity. While this identification strategy is reasonable for dealers, it is problematic for customers who presumably spend long periods of time out of the market: clearly, observing that a customer trades once a year does not imply that it takes a year to find a counterparty. Gavazza (2016) addresses this issue using

[4] See Bessembinder, Spatt, and Venkataraman (2020) for a survey

a structural model, taking advantage of aggregate information about the total number of real assets (in his case, aircraft) for sale at a time. Pintér and Üslü (2021), alternatively, use joint observation of trade size and frequency to indirectly identify search intensities. We propose a more direct approach, based on granular observations, which does not rely on the restrictions imposed by a specific structural model.

Finally, our approach is related to the large literature that attempts to estimate the key objects of interest in the standard sequential search model of McCall (1970), which was first used in financial economics by Garbade and Silber (1976). Early attempts to do so in a labor market context include Kiefer and Neumann (1979) and Flinn and Heckman (1982), among others. As in labor economics, this simple partial equilibrium model is a natural starting point for interpreting micro data, as it helps rationalize failed inquiries, repeated attempts to trade, and price dispersion.[5] However, while we find it useful to formulate a search-theoretic model to motivate our empirical exercise and interpret its findings, it's important to note that our measurement does not impose theoretical restrictions from the model.

2 Data

Our main source of data is MarketAxess (MKTX), the leading electronic trading platform in the corporate bond market. Prior to the introduction of MKTX, in 2000, the corporate bond market operated almost exclusively under a "voice-based" trading system, whereby customers would sequentially contact dealers (via telephone or chat) one at a time to solicit a quote. Stepping into this market, MKTX offered a trading platform allowing buy-side traders (henceforth customers) to query multiple dealers at once via an electronic request for quote (RFQ), thus reducing the time-consuming process of gathering quotes and potentially increasing competition across dealers.

[5]Naturally, understanding the process that generates dealers' offers requires expanding the model to include an explicit analysis of the market structure of the dealer sector, along with the optimal strategies of (potentially heterogeneous) dealers. Given the scope of the current paper, we leave this extended analysis for future work.

As of the third quarter of 2022, MKTX accounts for approximately 21% of total trading volume in the corporate bond market.[6]

When requesting a quote on the MKTX platform, customers specify the bond they wish to trade, the desired quantity, the trade direction or "side" (buy or sell), and the duration of the auction (usually between 5 and 20 minutes). Once submitted, a customer inquiry is sent to a list of pre-authorized dealers.[7] On the receiving end, dealers observe the details of the inquiry, including the customer's identity. The receiving dealers may respond to the inquiry with a quote, but are not obligated to do so. At the end of the auction, customers observe the terms of the replies (if any), and can choose to either accept one of the offers or pass.[8]

Our sample from MKTX covers all trading activity from January 3, 2017 to March 31, 2021. The data contain detailed information on customer inquiries, dealer responses, and customer trading decisions. More specifically, for each inquiry, we observe the submission time (stamped at the second), an anonymized customer identifier, the CUSIP (Committee on Uniform Securities Identification Procedures) number of the requested bond, the requested quantity, the trade side (buy or sell), the number of dealers who received the request, and several other attributes. For every response to an inquiry, we observe the anonymized identifier of the responding dealer together with his quote. For inquiries that result in a transaction, we observe the time at which trade occurs and the terms of trade. Note that we observe all inquiries, including those that do *not* result in a trade, either because the inquiry receives no responses or because the customer chooses to reject all responses.

Importantly, when an inquiry fails to trade on MKTX, a customer may trade outside the platform via voice. In the next section, we describe how we attempt to identify these trades using the enhanced version of the Trade Reporting Compliance Engine (TRACE) data set provided by

[6]Source: MarketAxess quarterly report for 2022Q3, available from: https://investor.marketaxess.com.

[7]Starting in 2012, MKTX initiated Open Trading, a trading protocol that enables all-to-all trading in the corporate bond market. This protocol allows other investors as well as non-pre-authorized dealers to respond to requests for quotes. Approximately 15% of MKTX auctions are won by responses submitted through Open Trading. For a comprehensive analysis of the Open Trading protocol, see Hendershott, Livdan, and Schürhoff (2021).

[8]The main variation in dealers' offers is price. In principle, dealers can respond to an offer with a different quantity, but in practice more than 97% of dealer responses are at the quantity level requested by the customer.

FINRA. The TRACE database contains detailed reports of every successful trade, whether it has an electronic or voice origin. When working with TRACE, we filter the data following the standard procedure laid out in Dick-Nielsen (2014). We merge the cleaned data set with the Mergent Fixed Income Securities Database (FISD) to obtain bond fundamental characteristics (e.g., credit ratings, amount outstanding, coupon rates, etc.) Following the bulk of the academic literature, we exclude variable-coupon, convertible, exchangeable, and puttable bonds, as well as asset-backed securities, privately placed instruments, and foreign securities. We also exclude primary market transactions.

Finally, we measure transaction costs as a markdown or markup relative to the benchmark provided by MKTX, called Composite+ (CP+).[9] CP+ is the proprietary algorithmic pricing engine for corporate bonds from MKTX. It is designed to provide an unbiased two-sided market forecast for institutional-size trades. The engine outputs reference bid and ask prices at a high frequency (every 15 to 60 seconds). These forecasts can be used to benchmark a significant fraction of TRACE records: 90% of high-yield TRACE records can be matched to a standing CP+ forecast; that figure goes up to 95% for investment-grade bonds.

The construction of the forecasts follows two steps. First, MKTX trains a machine learning (ML) algorithm using three distinct sources of bond trading data: (1) historical TRACE prints; (2) indicative bond price data streamed by dealers; and (3) request for quote responses sent by liquidity providers on the MKTX trading platform. Beyond trading data, MKTX uses bond level information and other broad market data, such as CDX levels, to train the prediction engine. The engine is recalibrated overnight at a daily frequency. Second, the calibrated engine is used over the next trading day to generate real-time reference bid and ask prices of individual bonds using all available intraday information.

2.1 The query process: parent and child orders

To give the reader a sense of how the query process works and to motivate the way we organize and analyze the data, we believe it is helpful to present some representative examples of inquiries.

[9]For more details about Composite+, see https://www.marketaxess.com/price/composite-plus.

First, panel (a) of Table 1 provides an example of a successful inquiry. In this example, a customer submitted an inquiry to buy $300,000 in par value of an investment-grade bond issued by Bank of America. The customer received six replies from dealers, whose anonymized identifiers are provided in column (6). Note that, because the bond in question is investment-grade, dealer responses in column (7) are in terms of yield spread relative to a benchmark Treasury bond (a higher yield spread implies a lower purchasing price). As we can see from this column, dealers' quoted yield spreads vary between 126.37 and 129.70 basis points. In the second row of column (9), the entry "Done" shows that the customer accepted the best (highest) offer.[10] In our sample, 67% of all inquiries result in a successful trade.[11]

[Table 1 about here.]

Panel (b) of Table 1 provides an example of unsuccessful inquiry. This inquiry was submitted by the same customer and for the same bond as the inquiry reported in panel (a), but this time, the customer requested to purchase an amount of $490,000 in par value instead of $300,000. A total of nine dealers responded to the customer's new request. By comparing the identifiers of responding dealers for both inquiries, we see that five of the six dealers who responded to the first inquiry also responded to the second untraded inquiry. Four additional dealers, who had not replied to the first inquiry, replied to the second inquiry. However, the customer decided to pass on the best offer (a yield spread of 127.01), as indicated by the "did not trade" (DNT) flag in the last column. In our sample, 16% of inquiries that receive at least one response do not trade. An additional 18% of inquiries do not receive any response.

While customer inquiries are informative about the trading process in and of themselves, a careful examination of the data reveals that individual inquiries are often parts of larger trading orders. As a result, individual inquiries should not always be treated as independent observations. To help the reader see why, Table 2 reports all the inquiries that the customer in our previous

[10] In the last row of column (9) in Table 1, the entry "Cover" identifies the second best offer. MKTX informs dealers who submit the second best offer of the rank of their quote. Dealers who submit lower-ranked offers do not learn their relative position in the auction.

[11] While we examine a different time period, we find a fraction of successful trade consistent with the findings of Hendershott and Madhavan (2015).

examples (Table 1) submitted to purchase this particular Bank of America bond over a six month period. To save space, we do not report the responses that each inquiry received, and report only whether or not a given inquiry resulted in a trade (see column 7). Note that the first and second inquiries reported in Table 2 correspond to the inquiries reported in panel (a) and panel (b) of Table 1.

Notice immediately that the customer made repeated *successful* purchase inquiries for the same bond over an eight day period. Of the six inquiries, four were successful and led to the purchase of 300, 490, 290, and 680 bonds (with $1,000 par value) for a total of 1760 bonds. This anecdotal evidence suggests that customers sometimes execute large orders by submitting a sequence of smaller inquiries. To give further credence to this interpretation, Figure 1 plots the daily number of purchase inquiries submitted by this customer over a six month horizon. The figure makes clear that the customer's inquiries over that horizon are concentrated in mid-August 2017, which supports the view that the individual inquiries are part of a larger order and not independent events.

The second noteworthy feature of Table 2 is that the customer twice followed an *unsuccessful* inquiry by resubmitting an identical inquiry (same bond, quantity, and trade side) soon afterward. This phenomenon is first observed after the second inquiry and again after the fourth. While both of these unsuccessful inquiries received multiple dealer responses, the customer chose to pass.[12] Hence, the example in Table 2 suggests that even when customers are able to simultaneously contact a large number of dealers, sequential search remains a feature of the trading process in the U.S. corporate bond market.

These patterns of trade are widespread. For example, about a third of trading volume can be attributed to a parent order with two or more child orders, and approximately a quarter of child orders have at least two inquiries. Hence, we argue that a natural first step is to organize RFQs into clusters, representing the total quantity of a particular bond that a customer is attempting to trade, which we refer to as the "parent" order. Within each parent order, we further partition the set of inquiries into sets of "child" orders in which the customer requests a specific quantity of the bond.

[12]For the second inquiry, this can be seen in panel (b) of Table 1. To save space, the responses associated with the fourth inquiry are not reported.

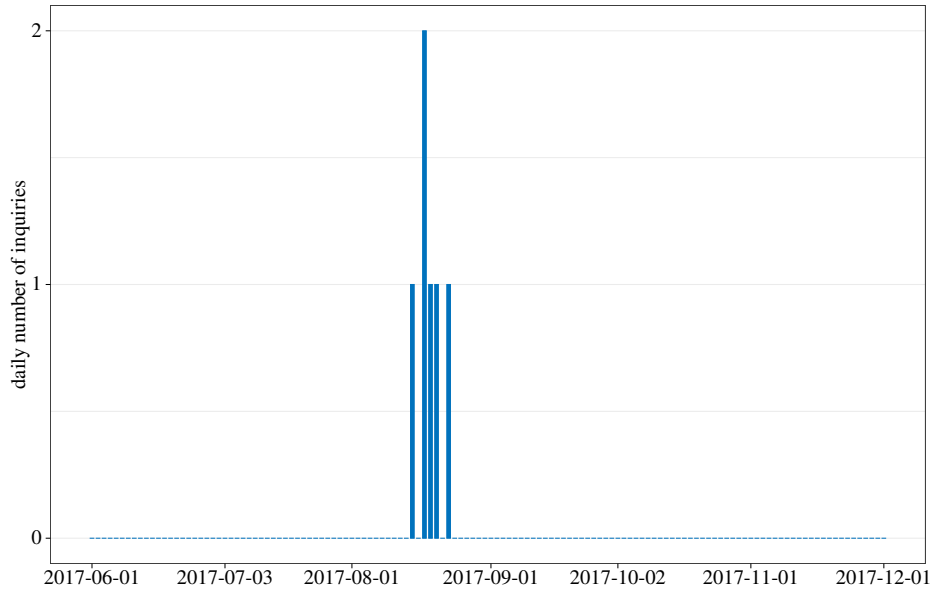


Figure 1. Inquiry cluster
This figure shows an inquiry cluster for a customer purchase for an 11-year, 3.824% investment-grade bond issued on January 17, 2017 by Bank of America over a six-month period in 2017.

We borrow the parent and child order terminology from the equity market literature on institutional trading where large (parent) orders are often split into smaller (child) orders for execution. In the example above, as one can see in columns (8) and (9) of Table 2, all six inquiries make up a single parent order—where the customer attempts to trade 1760 units of this particular bond over an eight day period—and this parent order is split into four smaller child orders.

[Table 2 about here.]

More precisely, since the data itself does not explicitly identify parent and child orders, we employ the following procedure to construct the appropriate identifiers. First, to construct parent orders, we group all inquiries made by a specific customer for a given bond and trade side until we do not observe a new inquiry with the same characteristics (customer, bond, trade side) for N_p days since the last inquiry. The time cutoff delimiting parent orders is admittedly arbitrary. In our main specification, we use a cutoff of five days. However, our main results are not sensitive to this choice; we obtain qualitatively similar results with a cutoff of ten days.

Second, we construct child orders by looking at repeated inquiries from a given customer for the same bond, the same trade side, *and the same requested quantity*. We consider all inquiries

with these characteristics as part of the same child order until either (i) the most recent inquiry of the child order led to an electronic trade on MKTX; (ii) the customer submitted a new inquiry requesting a different quantity, in which case we initiate a new child order with the updated quantity; or (iii) there is no new inquiry with the same characteristics (same customer, bond, trade side, and trade size) for more than N_c days, where $N_c \leq N_p$. When no new inquiry has been submitted for more than N_c days, we consider the execution of the child order unsuccessful on MKTX. Here again, the threshold N_c is arbitrary. While we use a cutoff of five days in our main specification, our main results are not sensitive to this choice.

There are two reasons why a child order may be unsuccessful on MKTX. First, the customer might have given up trading the bond. Second, the customer might have traded the bond via voice. These two outcomes have different economic implications and should be distinguished. Ideally, we would match customer inquiries on MKTX that result in a voice trade using the corresponding TRACE record. However, since TRACE does not report customer identities, it is impossible to match a child order that is traded via voice to its corresponding TRACE record with certainty. Fortunately, this issue can partially be overcome since most corporate bonds trade only a few times a day or less. As a result, the likelihood that two different customers would trade the same quantity of the same bond within a few days is arguably low. We thus infer the occurrence of a voice trade by verifying if there exists a record in TRACE with the same characteristics as the unsuccessful child order (same bond, traded quantity, trade side) within five days of that child order's last on MKTX. In the rare cases where there are multiple matches, we select the closest one in time.

2.2 Summary statistics

We could, in principle, conduct our analysis in two ways: at the level of parent orders, or at the level of child orders. However, the splitting of a parent order into child orders may be driven by considerations other than search, such as asymmetric information (as in, e.g., Kyle, 1985). For this reason, we find it more natural to study the sequential search process using child orders as our main unit of observation. We begin this section by presenting some summary statistics, explaining

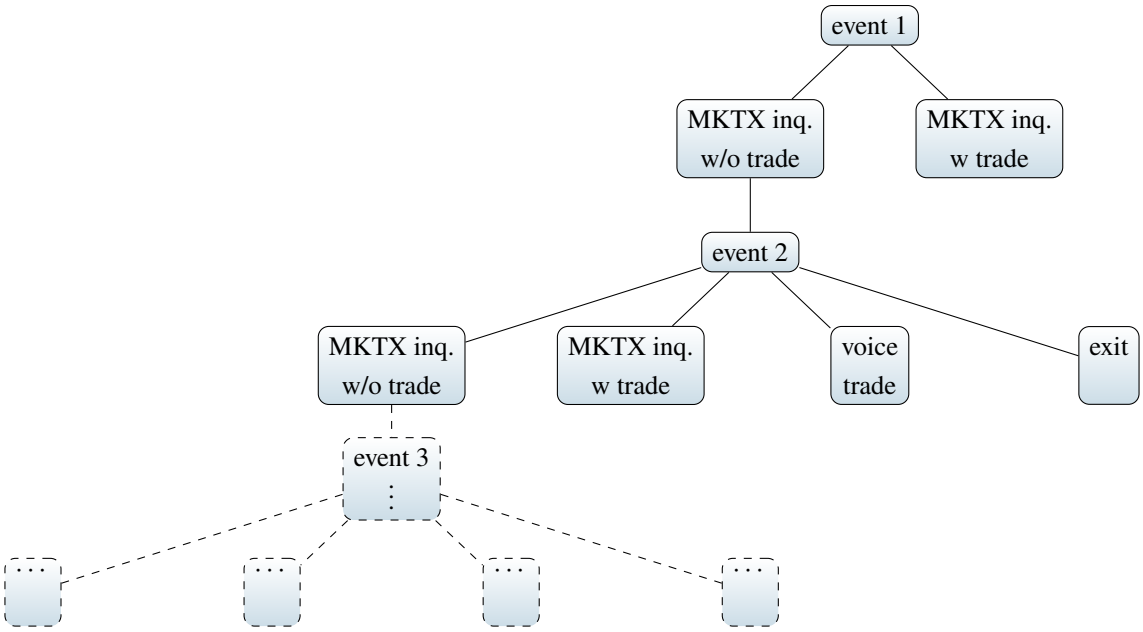


Figure 2. A child order event tree
A child order can be viewed as a sequence of events. Each element of the sequence is one of four possible events: a MKTX inquiry without trade, a MKTX inquiry with trade, a voice trade and, if the child order ends without a trade, an exit. By construction, the first event is always an inquiry on MKTX, either without or with trade.

how our sample differs from previous studies on the corporate bond market, and providing some preliminary evidence about the sequential trade process.

A child order is a sequence of events. Each element of the sequence is one of four possible events. First, the customer may make an inquiry on MKTX that fails to produce a trade. Second, the customer may make an inquiry on MKTX that results in a trade with one of the dealers that responded. Third, we may find that the customer traded the desired bond-quantity pair outside of the MKTX platform, via voice trade, within a short period of time. Fourth, the customer may give up on the trade and exit — either sending an inquiry for a different amount or abandoning the trade altogether. By construction, the first event in any child order that we observe is always an inquiry on MKTX, without or with trade. We can measure the time elapsed to the next event, unless it is an exit. Figure 2 illustrates a child order event tree.

Our focus on child order sets us apart from previous studies, such as Hendershott and Madhavan (2015) or O'Hara and Zhou (2021), who consider the universe of all inquiries and/or of all trades on MKTX. A simple way to illustrate the conceptual difference between child orders and inquiries

is to calculate trade probabilities. Since child orders include repeated inquiries on MKTX, they are naturally associated with a larger trading probability than inquiries alone. The difference is economically significant: in our sample, approximately 75% of child orders are traded on MKTX, while the trade probability at the inquiry level is 67%.

In Table 3 and Figure 3, we present the results of a logit regression. The dependent variable is whether trade occurs on MKTX and the independent variables are indicator functions for customer and trade characteristics. The "baseline category" is a round lot, investment-grade, buy request, for an above-median turnover and amount outstanding, and below-median time to maturity bond, from a well connected customer. Column (1) and (2) of Table 3 present estimates at the child order and inquiry levels, respectively. The unit of the coefficients is log odds ratio of trade. For example, the intercept in the column (1) shows that the odds ratio of trade for the baseline category is $\exp(2.462)$, leading to the probability of trade of $\exp(2.462)/(1+\exp(2.462)) = 92\%$. In other words, at the child order level, the probability of trade for the baseline category is 92%. At the inquiry level in column (2), the odds ratio is smaller by about 18 percentage points, corresponding to a trade probability of 88%. So, one can see that child orders are executed with higher probability than inquiries. In Figure 3, the blue bars represent inquiry-level trade probabilities, and the combined blue and grey bars represent child-order trade probabilities.

[Table 3 about here.]

The estimates for covariates in Table 3 are interesting as well. For example, sequential trade matters a great deal for the "least connected" customers, defined as those who elicit a relatively low number of responses from dealers. The trade probability at the inquiry level is about 45%, but it is about 55% at the child order level. One can also see that the probability of trade fell at both inquiry and child order levels during the COVID-19 crisis in March 2020, but that the fall was much less dramatic at the child order level: at the inquiry level, the trade probability falls to about 75%, but at the child order level it falls much less, to 83%. This suggests another way sequential trade matters: during stressful events, it is harder for customers to obtain good quotes on MKTX, but investors could compensate for it by waiting.

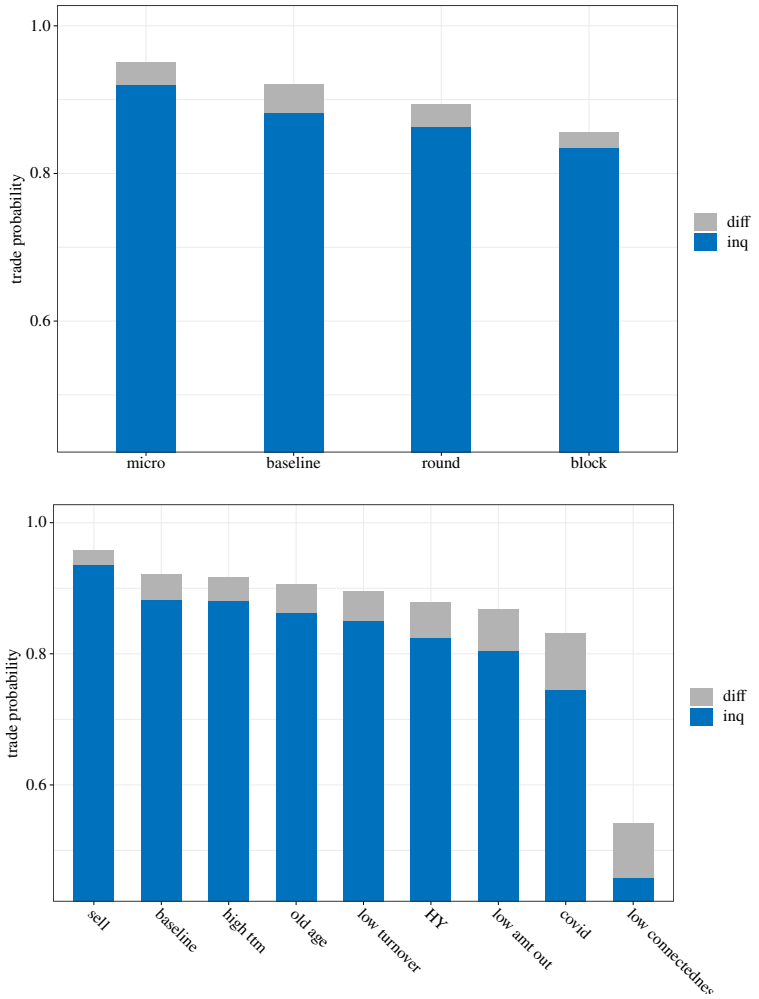


Figure 3. Estimated trade probability on MKTX at inquiry and child order levels
This figure compares the estimated trade probability on MKTX using logit regression estimates from Table 3. The blue bars present trade probabilities at the inquiry level. The gray bars shows the extra trade probability for a child order, taking into account the option to make repeat inquiries. The top panel shows trade probabilities for different size categories. The bottom panel presents trade probabilities for non-size categories. Indicators for size and non-size categories are defined in Table 8. The baseline category is an odd-lot purchase of an investment-grade bond, with high turnover, low time-to-maturity, and high amount outstanding, during normal times, for a connected investor.

One may wonder whether our sample significantly differs from its inquiry- or trade-level counterparts in other dimensions as well.[13] Table 4 shows that this is not the case: child-order and inquiry-level summary statistics are broadly the same for trade direction and size and bond characteristics. As in previous studies, we find that trade sizes on MKTX are smaller and bond credit risk is lower than in the market at large. To measure the inter-arrival time of trading opportunities,

[13]For example, suppose that high-yield bonds trade after twice as many inquiries as investment-grade bonds. Then we would find that the number of high-yield inquiries is twice that of high-yield child orders.

we will need to restrict the sample further to child orders with at least one failed inquiry. Column (3) of Table 4 shows that the summary statistics remain similar, though the sample is now more selected towards high-yield bonds and inquiries of larger size since both are less likely to trade at the first MKTX inquiry.

[Table 4 about here.]

Table 5 offers some descriptive statistics about child orders. The first row shows that the probability that a child order does not trade at the first inquiry is nontrivial, about 0.28. The following rows provide the frequency distribution over the next event in the child order, conditional on the number of failed inquiries to date. For example, the second row shows that if the first inquiry fails, the probability that the following event is a failed inquiry on MKTX is 0.17, the probability that there is a successful inquiry on MKTX is 0.09, the probability that there is a voice trade is 0.24, and the probability of an exit is 0.49. We will argue later that these probabilities are not obvious to interpret because of competing risk and selection biases. Notwithstanding these issues, there are a few takeaways. First, trade is sequential: the probability of failing an inquiry is nontrivial, and customers often submit repeat inquiries. Second, trade is non-exclusive: if the first inquiry fails, the child order may eventually trade on MKTX or voice.[14] The third takeaway is that exit is nontrivial: the probability that a child order ends without trade is large. Fourth, the summary statistics in Table 5 show that the frequency distribution over the four events depends on the number of failed inquiries – a form of duration dependence.

[Table 5 about here.]

Table 6 presents inter-arrival times between events in child orders. For example, after one failed inquiry, the average time to the next traded inquiry on MKTX is 0.65 business days. However, as we argue below, this estimate is clearly biased downwards, since observing this event requires that none of the other events occur first.

[14] While the probability of a voice trade is larger than that of an MKTX trade, the ratio is not as large as the relative volume of voice to MKTX volume. This suggests that, although trade is nonexclusive, the customers in our sample are using MKTX more intensely than the general population.

[Table 6 about here.]

3 The sequential trade process: theory and evidence

3.1 A McCall (1970) model of a child order

In this section, we formulate and solve a sequential trade model of a child order in the style of McCall (1970), which was first applied to financial markets by Garbade and Silber (1976). This theoretical detour serves two purposes. First, it is a simple and natural theoretical framework to interpret our child order data; in particular, it helps clarify competing risk and selection biases in child order statistics and motivates the statistical models we estimate later. Second, since the McCall (1970) model is the workhorse, partial equilibrium model of sequential search, it constitutes a key building block in virtually all search-based models of OTC markets. Interpreting our empirical evidence through the lens of this sequential search model offers guidance for the quantitative values of key parameters—allowing us to infer, e.g., the time it takes to trade—and highlights which dimensions of the model fit the data well, and which dimensions must be enriched in order to match certain features of the data.

The model. Time is indexed by $t \in [0, \infty)$. We consider a child order to sell a perpetual par bond, that is, a perpetuity with a coupon rate that is equal to the interest rate, r. We assume that the seller is risk-neutral with discount rate r and is distressed, in that she values the bond below its par value of 1. Specifically, when she holds the bond, she derives a flow utility $r - c$, for some distress cost $c > 0$. The seller recovers from distress with intensity γ. Upon recovering, we assume that the seller's continuation value is equal to the par value of the bond, she stops searching, and exits the market.[15] We focus here on a customer looking to sell, for simplicity, but the analysis of a purchase is symmetric.

Consistent with the child order tree of Figure 2, we take $t = 0$ to represent the time at which

[15] We discuss alternative assumptions after Proposition 1.

the seller makes her first inquiry on the electronic market.[16] If the first inquiry is unsuccessful, the seller makes inquiries on the electronic or the voice market with Poisson intensities λ_e and λ_v, respectively. At this level of abstraction, the arrival rate of trading opportunities could represent frictions that derive from either side of the market. For example, the source of the friction could be that the customer simply can't find a dealer willing to buy the asset at an acceptable price, as in the literature following Duffie, Gârleanu, and Pedersen (2005). Alternatively, it could be that the customer is busy with other tasks and not actively trading in the market at all times, as in Biais and Weill (2009) and Biais, Hombert, and Weill (2014).

After an inquiry in the electronic market, the seller receives $j \in \{0, 1, 2, \ldots\}$ offers with probability q_j. We represent an offer as a bid $1 - m$, where m is the markdown over the bond par value of 1. We assume further that each offered markdown is drawn independently according to the cumulative distribution function (CDF) $G_e(m)$. Correspondingly, when she makes an inquiry in the voice market, the seller receives just one offer, drawn according to the CDF $G_v(m)$. For simplicity we assume that, for both distributions, the lower bound of the support is 0. As will be clear below, the optimal trading strategy of the seller depends on two "sufficient statistics." First the *total* Poisson intensity of inquiries,

$$\lambda = \lambda_e + \lambda_v,$$

and, second, the CDF over the *lowest* markdown, conditional on an inquiry,

$$F(m) = \frac{\lambda_e}{\lambda_e + \lambda_v} \sum_{j=0}^{\infty} q_j \left[1 - (1 - G_e(m))^j\right] + \frac{\lambda_v}{\lambda_e + \lambda_v} G_v(m).$$

The first term in this equation is the probability of making an inquiry on the electronic market, multiplied by the probability that the smallest markdown among j offers is less than m. The second term has the same interpretation, but for the voice market.

Given this notation, the Hamilton Jacobi Bellman (HJB) equation for the seller's value at any

[16] Alternatively, one may interpret $t = 0$ as the time at which the seller becomes distressed.

time $t > 0$ is

$$rV = r - c + \lambda \int \max\{1 - m - V, 0\}\, dF(m) + \gamma(1 - V). \tag{1}$$

The first term on the right-hand side is the flow value of holding the asset, i.e., the coupon r net of the distress cost c. The second term is the option value of search: the seller makes an inquiry with intensity λ, her best offer is distributed according to $F(m)$, and she accepts if the price $1 - m$ is larger than the value of continuing search, V. The third and last term is the expected flow utility if the seller recovers/exits. As is standard, the HJB shows that the optimal trading strategy of the seller is entirely characterized by the reservation markdown

$$m^\star \equiv 1 - V. \tag{2}$$

That is, when she makes an inquiry, the seller trades if and only if the lowest markdown she receives is less than $m^\star$. To obtain an equation for $m^\star$, we substitute (2) in the HJB and obtain, after integration by parts, our version of McCall's celebrated equation, summarized in the following proposition.

Proposition 1 *The reservation markdown of a seller is the unique solution to*

$$m^\star = \frac{c}{r + \gamma} - \frac{\lambda}{r + \gamma} \int_0^{m^\star} F(m)\, dm. \tag{3}$$

The reservation markdown $m^\star$ increases with the distress cost c, decreases with the interest rate, r, decreases with the exit rate, γ, decreases with the inquiry intensity, λ, and increases in response to a first-order stochastic dominance shift in the distribution of the best markdown, $F(m)$.

The first term in Equation (3), $c/(r + \gamma)$, is the expected present value of the seller's distress cost. It represents the monopsony markdown: the maximum markdown a seller would be willing to accept if she received just one take-it-or-leave-it offer by a dealer, and no offer forever after. The

optimal reservation markdown is less than the monopsony markdown because of the option value of searching for another offer.

The comparative statics for the reservation markdown are similar to those obtained in the classical job-search setting, with the exception of the effect of varying $r + \gamma$. The reason is that, in our setting, increasing $r + \gamma$ impacts the seller's problem in two ways. First, as in job-search models, increasing $r + \gamma$ reduces the option value of search which, all else equal, increases the reservation markdown. Second, and new to this setting, it decreases the present value of the seller's distress costs, which decreases the reservation markdown. The second effect, it turns out, always dominates in our setting.[17]

Next, we use this simple model as an aid to interpret our child order data. Recall the child order tree of Figure 2, where a child order is viewed as a sequence of events. According to the model, there is a new event in the child order tree with intensity

$$\lambda_e + \lambda_v G_v(m^\star) + \gamma.$$

Conditional on an arrival, the new event is drawn independently from the arrival time according to the following distribution. The new event is an inquiry without trade on the electronic market with probability

$$\pi_1 = \frac{\lambda_e \sum_{j=0}^{\infty} q_j (1 - G_e(m^\star))^j}{\lambda_e + \lambda_v G_v(m^\star) + \gamma},$$

[17]For now, we have assumed that the seller exits the market when she recovers from distress: as shown in the HJB equation (1), her continuation value is set to the par value of the bond (1) upon recovery. But one may consider other plausible assumptions: for example, an exit in our data could occur because the seller goes back to the market with a different inquiry, e.g., for another quantity or a closely substitutable bond. In the analysis of the McCall (1970) model above, this amounts to changing the continuation value. Assume, for example, that when an exit occurs, the seller submits another child order for almost the same quantity or a nearly identical bond. Then, in the HJB equation, the continuation value is V instead of 1, and the reservation markdown equation is almost the same: the appropriate discount rate is now r instead of $r + \gamma$.

it is an inquiry with trade on the electronic market with probability

$$\pi_2 = \frac{\lambda_e \sum_{j=0}^{\infty} q_j \left[1 - (1 - G_e(m^\star))^j\right]}{\lambda_e + \lambda_v G_v(m^\star) + \gamma},$$

it is a trade on the voice market with probability

$$\pi_3 = \frac{\lambda_v G_v(m^\star)}{\lambda_e + \lambda_v G_v(m^\star) + \gamma},$$

and it is an exit with probability $\pi_4 = 1 - \pi_1 - \pi_2 - \pi_3$.

The formulae above illustrate two sources of bias that make interpreting child order statistics difficult. We discuss these two sources of bias below.

Competing risk bias. First, since the event type is drawn independently from the event arrival time, it follows that the *observed* expected arrival time of any of the four events is given by

$$\overline{\tau} = \frac{1}{\lambda_e + \lambda_v G_v(m^\star) + \gamma}.$$

Notice that this observed expected arrival time is *lower* than the actual arrival time of the event. For example, the actual arrival time of a voice trade is $1/(\lambda_v G(m^\star))$. This is a classical survivor bias induced by competing risk (e.g., Flinn and Heckman, 1982; Katz and Meyer, 1990; Honoré and Lleras-Muney, 2006) created by the arrival of other events. Imagine for example, that sellers exit the market very fast. Then the only trades on the voice market we would observe are those that occur sufficiently quickly, before an exit shock.

The formulae above show that there is a simple way to correct for this survivor bias. For example, the true expected time to trade on voice is equal to the ratio $\overline{\tau}/\pi_3$. As we will show below, this correction can be made more generally using a Maximum Likelihood approach, conditional on observable child-order characteristics.

Selection bias. In the data, we can control for several observed characteristics of child orders, such as bond type, trade size, and a natural measure of customer connectedness. But there may be other characteristics that are difficult to control for based on observables, including the distressed cost of a seller, c; her inquiry intensities, λ_e or λ_v; her ability to elicit responses from dealers, $\{q_j\}$; or her exit intensity, γ. Such unobserved characteristics create classical selection issues that could explain the apparent dependence of event probabilities on the number of failed inquiries, shown in Table 5.

To fix ideas formally, suppose that heterogeneity in child orders can be summarized by a one-dimensional type variable $x \in [\underline{x}, \overline{x}]$. This specification allows for heterogeneity in all structural variables (λ_e, λ_v, γ, c, r, G_e, G_v, and so on), provided that there is a fixed relationship between the type x and each structural variable in the cross-section of child orders. Then, the measure of type-x child orders with $n \geq 1$ failed inquiries, $\mu(x \mid n)$, satisfies the following inflow-outflow equation:

$$\lambda_e(x) \left[\sum_j q_j(x) \left[1 - G_e(m^\star(x) \mid x) \right]^j \right] d\mu(x \mid n-1) \\ = \left[\lambda_e(x) + \lambda_v(x) G_v(m^\star(x) \mid x) + \gamma(x) \right] d\mu(x \mid n).$$

The left-hand side is the inflow generated by child orders that make unsuccessful inquiries on the trading platform. Similarly, the right-hand side is the outflow generated by child orders that make inquiries on the trading platform, trade on the voice market, or exit. Taken together, these inflow-outflow equations imply that

$$d\mu(x \mid n) = \pi_1(x)^n \, d\mu(x \mid 0), \quad \text{where} \quad \pi_1(x) \equiv \frac{\lambda_e(x) \left(\sum_j q_j \left[1 - G_e(m^\star(x) \mid x) \right]^j \right)}{\lambda_e(x) + \lambda_v(x) G_v(m^\star(x) \mid x) + \gamma(x)}. \quad (4)$$

According to (4), the measure of type-x child orders with n failed inquiries declines geometrically with n. As discussed above, the geometric coefficient, $\pi_1(x)$, is simply the probability that a type-x inquiry on the electronic trading platform fails to trade (the left-most branch of event 2 in the child-order tree of Figure 2).

Next we show that the direction of the selection bias depends on the geometric coefficient, $\pi_1(x)$. Namely, let

$$dH(x \mid n) = \frac{d\mu(x \mid n)}{\int_{\underline{x}}^{\overline{x}} d\mu(y \mid n)},$$

denote probability distribution of x across child orders conditional on n failed inquiries. The following Lemma reports a key property of this distribution.

Lemma 1 *If $\pi_1(x)$ is an increasing (decreasing) function, then $H(x \mid n)$ first-order stochastically dominates (is first-order stochastically dominated by) $H(x \mid n-1)$.*

Lemma 1 shows that as the number of failed inquiries, n, increases, the sample of child order becomes more selected towards those investors who, in their child order tree, fail inquiries on the trading platform with higher probability. As a result, if x is unobservable to the econometrician, any outcome variable which is monotonically related to x will appear to be monotonically related to the number of failed inquiry.

For example, suppose child orders differ in terms of the customer's distress cost, c, but are otherwise identical. Then $\pi_1(c)$ is decreasing in c since more distressed sellers have a higher reservation markdown, $m^\star$. As a result, as the number of failed inquiries increases, the sample gets more and more selected towards less distressed customers. It follows that we should observe two key outcome variables, the trading probability and the transaction markdown, decline with the number of failed inquiries n.

3.2 Evidence about time to trade

We propose below a statistical framework to measure the time it takes customers to trade after their first inquiry on MKTX, correcting for the competing risk bias discussed above, and controlling for observable trade characteristics.

Maximum Likelihood Estimation. Our unit of observation i is an event node in the child order tree of Figure 2: specifically, the type and time of the event that follows an unsuccessful inquiry on MKTX. We index the $K = 4$ possible events by $k \in \{1, \ldots, K\}$. Event $k = 1$ is an inquiry on MKTX without trade, $k = 2$ is an inquiry on MKTX with trade, $k = 3$ is a voice trade, and $k = 4$ is an exit. We assume further that these events arrive at independent exponential times with intensity $\lambda(\theta_k' x_i) = \exp(\theta_k' x_i)$, where x_i is a vector of covariates for that child order. These covariates include trade size, bond characteristics, customers' characteristics, *and* the number of failed inquiries on MKTX; the latter is particularly important, in that it allows us to identify potential duration dependence.

Given this statistical framework, conditional on x_i, the event k occurs at time $\tau_i = t$ with probability density

$$\mathbb{P}\left(\tau_i = t, \omega_i = k \mid x_i\right) = \lambda(\theta_k' x_i) e^{-\sum_\ell \lambda(\theta_\ell' x_i) t}.$$

This formula is the product of the probability that event k occurs at time t, $\lambda(\theta_k' x_i) e^{-\lambda(\theta_k' x_i) t}$, and the probability that all other events, $\ell \neq k$, occur *after* time t, $e^{-\sum_{\ell \neq k} \lambda(\theta_\ell' x_i) t}$. This is the sense in which there are "competing risks": the probability density accounts for the fact that we observe event k only if the other events $\ell \neq k$ have not occurred before. Aggregating across events and the number of inquiries, the likelihood function is, evidently:

$$\prod_{i=1}^{n} \left(\sum_k \mathbb{I}_{\{\omega_i = k\}} \lambda(\theta_k' x_i) e^{-\sum_\ell \lambda(\theta_\ell' x_i) \tau_i} \right).$$

Recall that we never observe the time of an exit in our data set; rather, we observe only whether or not an exit occurred. Therefore, integrating with respect to τ_i when $\omega_i = K$, we obtain the likelihood for our actual observations:

$$\prod_{i=1}^{n} \left(\sum_{k \neq K} \mathbb{I}_{\{\omega_i = k\}} \lambda(\theta_k' x_i) e^{-\sum_\ell \lambda(\theta_\ell' x_i) \tau_i} + \mathbb{I}_{\{\omega_i = K\}} \frac{\lambda_K(\theta_K' x_i)}{\sum_\ell \lambda(\theta_\ell' x_i)} \right).$$

Taking logs, after a few lines of algebra, we obtain that the log-likelihood is $\sum_i L(\omega_i, \tau_i, x_i, \theta)$, where:

$$L_i(\omega_i, \tau_i, x_i, \theta) = \sum_k \mathbb{I}_{\{\omega_i=k\}} \theta_k' x_i - \mathbb{I}_{\{\omega_i \neq K\}} \left(\sum_\ell \exp(\theta_\ell' x_i) \right) \tau_i - \mathbb{I}_{\{\omega_i=K\}} \log \left(\sum_\ell \exp(\theta_\ell' x_i) \right).$$

We first gain some qualitative and quantitative intuition by deriving the *unconditional* Maximum Likelihood Estimator (MLE), i.e., the special case in which the only control is a constant.

Lemma 2 *Let $\hat{\pi}_k$ denote the empirical frequency of event k and $\hat{\tau}$ the empirical average inter-arrival time of an event $k \neq K$. Then, the MLE of θ_k is $\hat{\theta}_k = \log(\hat{\pi}_k / \hat{\tau})$.*

This is the same estimate that we intuitively derived in the previous section, when discussing the competing risk bias. Indeed, after a failed inquiry, the expected arrival time of any event is $\overline{\tau} = 1/(\sum_\ell \lambda_\ell)$, and the probability of event k is $\pi_k = \lambda_k / \sum_\ell \lambda_\ell$. This shows that $\lambda_k = \pi_k / \overline{\tau}$ and $\theta_k = \log(\lambda_k)$, which is the population counterpart of the estimator in Lemma 2. The estimation results are shown in Table 7.

[Table 7 about here.]

The results offer some guidance about the orders of magnitude of arrival times for different events. For example, the unconditional intensity of a voice trade is $e^{-3.40} = 0.0333$ per business hour, corresponding to an average time of $1/0.0333 = 29.96$ business hours, or about 3.3 business days (assuming 9 hours of trading per day). Importantly, the estimates clearly show that competing risk creates a significant bias in calculating time to trade: indeed, 3.3 business days is much larger than the observed average inter-arrival times shown in Table 6 above.

Next, we move to the *conditional* MLE, with controls for trade characteristics (coefficients shown in Table 8) and for the number of failed inquiries in the child order to date (coefficients shown in Table 9). All controls are dummies. The baseline category is an odd-lot purchase of an investment-grade bond, with high turnover, during normal times, for a connected investor, after one failed inquiry. There is no closed form solution for the estimators. However, since the likelihood

function is concave in the vector of coefficients $\theta = (\theta_k)_{1 \leq k \leq K}$, it can be maximized reliably using existing optimization packages.

Table 8 shows the manner in which the intensities of each event, $\lambda(\theta_k' x)$, vary with trade characteristics. The intensities for the baseline category are obtained by taking the exponential of the intercept. The marginal effect of other trade characteristics is given by the exponential of their respective coefficient. In particular, when the coefficient is sufficiently small, it approximates the marginal effect in percentage term: e.g., from the fourth row in column (2) of Table 8, the intensity of trade with MKTX for a bond rated Ca to C is approximately $-(e^{-0.4} - 1) \simeq 33\%$ lower than for an investment-grade bond.

The estimates in Table 8 demonstrate that intensities vary significantly with trade characteristics. Consider, for example, trade size. We observe that the intensity of trade with MKTX for micro size trades (with size < \$100,000) is larger than for odd lot trades (our baseline category with size between \$100,000 and \$1 million). The intensity for odd lots is larger than for round lots (with size between \$1 and \$5 million), which is larger than for block trades (with size larger than \$5 million). Interestingly the intensity of voice trade is not monotonic in trade size: for example, block trades trade faster on voice. Bonds with low turnover, and high-yield bonds, also have lower trading intensity, both on MKTX and the voice market. Interestingly, sales and purchases are asymmetric: customers trade faster when they sell, on average, than when they buy.

The last rows of Table 8 show the impact of customer connectedness on MKTX. To derive a measure of customer connectedness, we first regress the average number of dealer responses elicited by a particular customer on several control variables, including the customer's average inquiry size, the fraction of his requests that were sell vs. buy, and the fraction of requests that were for investment-grade vs high-yield bonds. We then rank customers into deciles based on residuals of this regression. This measure aims to proxy for customers' existing relationships with dealers or other unobserved characteristics of connected clients. We find that this measure of connectedness creates significant differences in trading intensity on MKTX. This finding is intuitive because a more connected customer receives more offers on average and so is more likely to obtain one that

falls below her reservation markdown. Finally, the fifth row of Table 8 reveals that the COVID-19 crisis (identified by inquiries submitted in March 2020) had a significant negative impact on the trading intensity.

[Table 8 about here.]

Table 9 shows that, after controlling for trade characteristics, the number of failed inquiries retains predictive power for the intensity of each event. The intensity of an inquiry on MKTX that doesn't result in trade increases with the number of failed inquiries. In contrast, the intensity of successful inquiries—i.e., inquiries on either MKTX or via voice that result in trade—decreases as the number of failed inquiries increases. Through the lens of the McCall (1970) model outlined in the previous section, this evidence suggests a role for unobserved child order characteristics, such as heterogeneity in distress cost or the arrival rate of trading opportunities.

[Table 9 about here.]

Time to trade. We define time to trade as the expected time a customer takes to trade, either on MKTX or on voice, if she is not subject to exit shocks. That is, we study a hypothetical world in which the investor never exits in the child order tree—say, because she continues to search when she receives an exit shock.

If the intensities did not depend on the number of failed inquiries, calculating time to trade would be simple. For example, from the intercepts in columns (2) and (3) in Table 8 or 9, the time to trade for our baseline category would be $1/(e^{-3.65} + e^{-3.38}) \simeq 16.65$ business hours, or 1.8 business days. However, the dependence of intensities on the number of failed inquiries requires us to modify this simple formula.

Formally, consider a child order after n failed inquiries. With a slight abuse of notation, let x_n denote the corresponding vector of covariates, where n stands for the number of failed inquiries to

date. Then, the expected time to trade satisfies the following recursive formula:

$$\begin{aligned} T(x_n) = \mathbb{E}\left[\tau' \mid x_n\right] &+ \mathbb{P}\left[\omega' = 1 \mid x_n\right] \times T(x_{n+1}) \\ &+ \mathbb{P}\left[\omega' = 2 \mid x_n\right] \times 0 \\ &+ \mathbb{P}\left[\omega' = 3 \mid x_n\right] \times 0 \\ &+ \mathbb{P}\left[\omega' = 4 \mid x_n\right] \times T(x_n). \end{aligned}$$

The first term is the expected time to the next event. The other terms add up to the expected continuation time to trade after the next event. Specifically, if the next event is an unsuccessful inquiry on MKTX, $\omega' = 1$, then there is one additional failed inquiry and the continuation time to trade is $T(x_{n+1})$. If the next event is $\omega' = 2$ or $\omega' = 3$, then trade occurs so the continuation time to trade is zero. The last line corrects the bias induced by the competing risk of exit: specifically, if the next event is an exit ($\omega' = 4$), we assume that the investor continues to search for a trade instead of exiting, so the continuation time to trade is $T(x_n)$.

Bringing the last term from the right-hand to the left-hand side, and using the exponential formula for expected inter-arrival time and event probability, we obtain the following recursion:

$$T(x_n) = \frac{1}{\lambda(\theta_1' x_n) + \lambda(\theta_2' x_n) + \lambda(\theta_3' x_n)} + \frac{\lambda(\theta_1' x_n)}{\lambda(\theta_1' x_n) + \lambda(\theta_2' x_n) + \lambda(\theta_3' x_n)} T(x_{n+1}). \tag{5}$$

We can use this formula to calculate the time to trade. Moreover, differentiating (5) with respect to x, we obtain a corresponding recursive formula for the gradient of time to trade, which allows us to apply the Delta method and obtain standard errors for the time to trade estimates. We illustrate our results in a sequence of figures, where we plot the expected time to trade, conditional on the number of failed inquiries and specific trade characteristics using estimates from the MLE. We represent the 95% confidence intervals by shaded areas surrounding the conditional expectation.

Figure 4 shows that, for our baseline category, the time to trade increases from about two trading days after one failed inquiry to nearly four trading days after ten failed inquiries. High-yield, old,

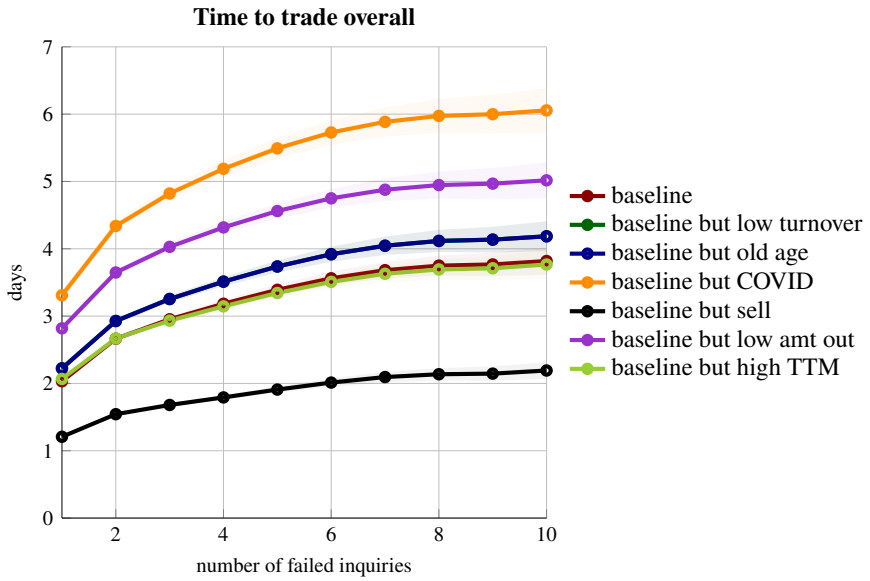


Figure 4. Estimated conditional time to trade from the MLE: observed trade characteristics
This figure plots the estimated time to trade from Equation (5), conditional on the number of failed inquiries and on observed trade characteristics except trade size and customer connectedness categories. "Sell" takes the value of 1 for a sale request, and zero otherwise; "COVID" takes the value of 1 if the RFQ is submitted in March 2020, and zero otherwise; "old age" takes the value of 1 if the bond's age is above the 75th percentile of the distribution, and zero otherwise; "low turnover" takes the value of 1 if the bond's quarterly turnover is below median, and zero otherwise; "high TTM" takes the value of 1 if the bond's time to maturity is above the sample median, and zero otherwise; "low amt out" takes the value of 1 if the bond's amount outstanding is below the sample median, and zero otherwise. The baseline category is an odd-lot purchase of an investment-grade bond, with high turnover, during normal times, for a connected investor, after one failed inquiry.

and low turnover bonds have a longer time to trade, though the difference is small relative to other covariates.

In Figure 5, we study the impact of trade size on time to trade. We observe that smaller trades are faster on MKTX. For example, after one failed inquiry, it takes 1.5 days to trade a micro-size bond, while the time it takes to trade a block-size inquiry is almost twice as long. This evidence is consistent with prior studies that show electronic trading is concentrated on smaller trades (e.g., Hendershott and Madhavan, 2015; O'Hara and Zhou, 2021).

Figure 6 shows that less connected customers, classified as customers that receive fewer offers from dealers, trade much slower on MKTX. For example, in the baseline category, the most connected customers (in the tenth decile of connectedness) trade after approximately 2.5 days

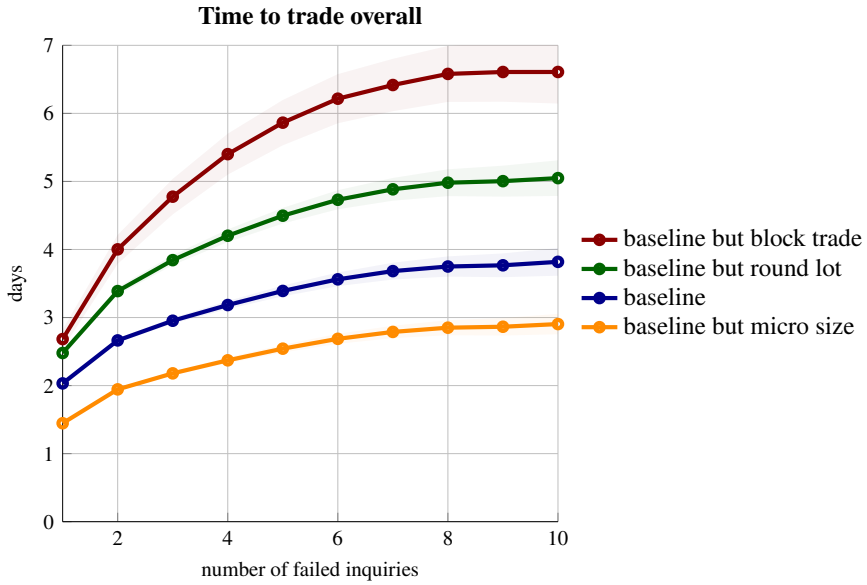


Figure 5. Estimated conditional time to trade from the MLE: impact of size
This figure plots the estimated time to trade from Equation (5), conditional on the number of failed inquiries and on trade size categories, and controlling for other observed trade characteristics. "Micro size" takes the value of 1 if the quantity of dealer response is below $100,000, and zero otherwise; "odd lot" takes the value of 1 if the quantity of dealer response is between $100,000 and $1 million, and zero otherwise; "round lot" takes the value of 1 if the quantity of dealer response is between $1 million and $5 million, and zero otherwise; "block trade" takes the value of 1 if the quantity of dealer response exceeds $5 million, and zero otherwise. The baseline category is an odd-lot purchase of an investment-grade bond, with high turnover, during normal times, for a connected investor, after one failed inquiry.

following two failed inquiries. For the least connected customers, in deciles 1 to 7, it takes almost 3.5 times more to trade on MKTX.

In Figure 7, we compare time to trade on MKTX to the one on voice for different trade size categories. The first takeaway is that, except for block trades, child orders trade much faster on MKTX than voice. This finding may be explained by the fact that customers initiate their first inquiries on MKTX and prefer to trade on the electronic platform, possibly for its execution quality rather than price discovery. Next, micro-size trades are faster than odd and round lots in both MKTX and the voice market, but block trades are much slower on MKTX. Again, this is not surprising, since, as mentioned above, smaller trades are more likely to be traded on electronic platforms.

It is important to keep in mind that our measurements are only descriptive. For example, the time to trade is presumably an endogenous outcome resulting from choices made by both sides of the market. The fact that customer purchases have longer time to trade could either indicate that

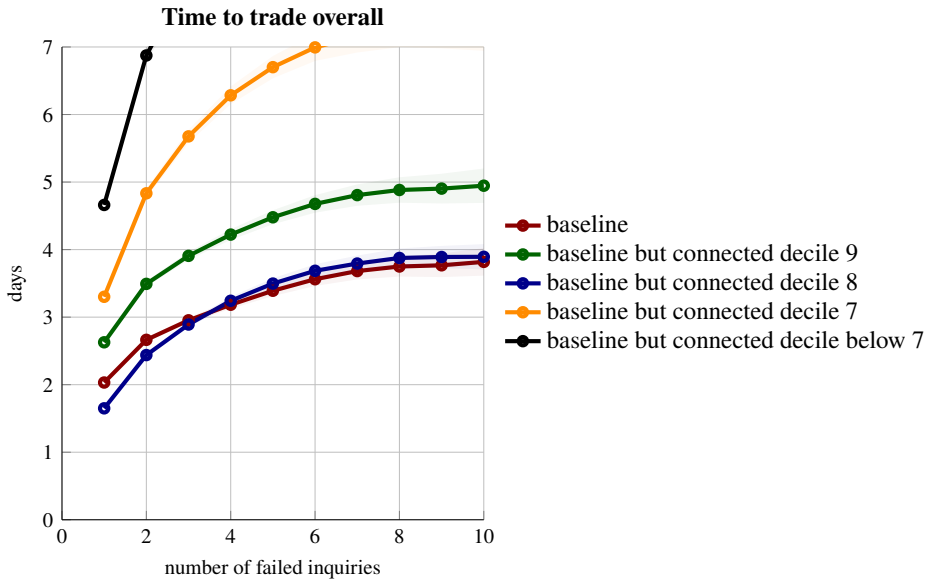


Figure 6. Estimated conditional time to trade from the MLE: impact of customer connectedness
This figure plots the estimated time to trade from Equation (5), conditional on the number of failed inquiries and on customer connectedness categories, and controlling for other observed trade characteristics. We first regress the average number of dealer responses elicited by a particular customer, on that customer's average inquiry size and fractions of requests for sell trades and high-yield bonds. We then rank customers into deciles based on residuals of this regression. "Connected decile 9" is an indicator for the customer being in decile 9, and similarly for other "Connected" indicators. The baseline category is an odd-lot purchase of an investment-grade bond, with high turnover, during normal times, for a connected investor (in decile 10), after one failed inquiry.

it takes time for dealers to source or locate a bond, or it could indicate that buyers are less eager to trade than sellers, so more willing to continue searching for a better price. The fact that time to trade increased during COVID could indicate that dealers were reluctant to accumulate inventories and changed their bidding behavior as a result.

3.3 The dependence of outcomes on the number of failed inquiries

We have found above that, after controlling for observed trade characteristics, the intensities estimated via MLE continue to depend on the number of failed inquiries. Figure 8 shows that this is true for other outcome variables as well. Panel (a) plots the trading probability on MKTX, calculated based on the MLE:

$$\frac{\lambda(\theta_2' x)}{\lambda(\theta_1' x) + \lambda(\theta_2' x)}.$$

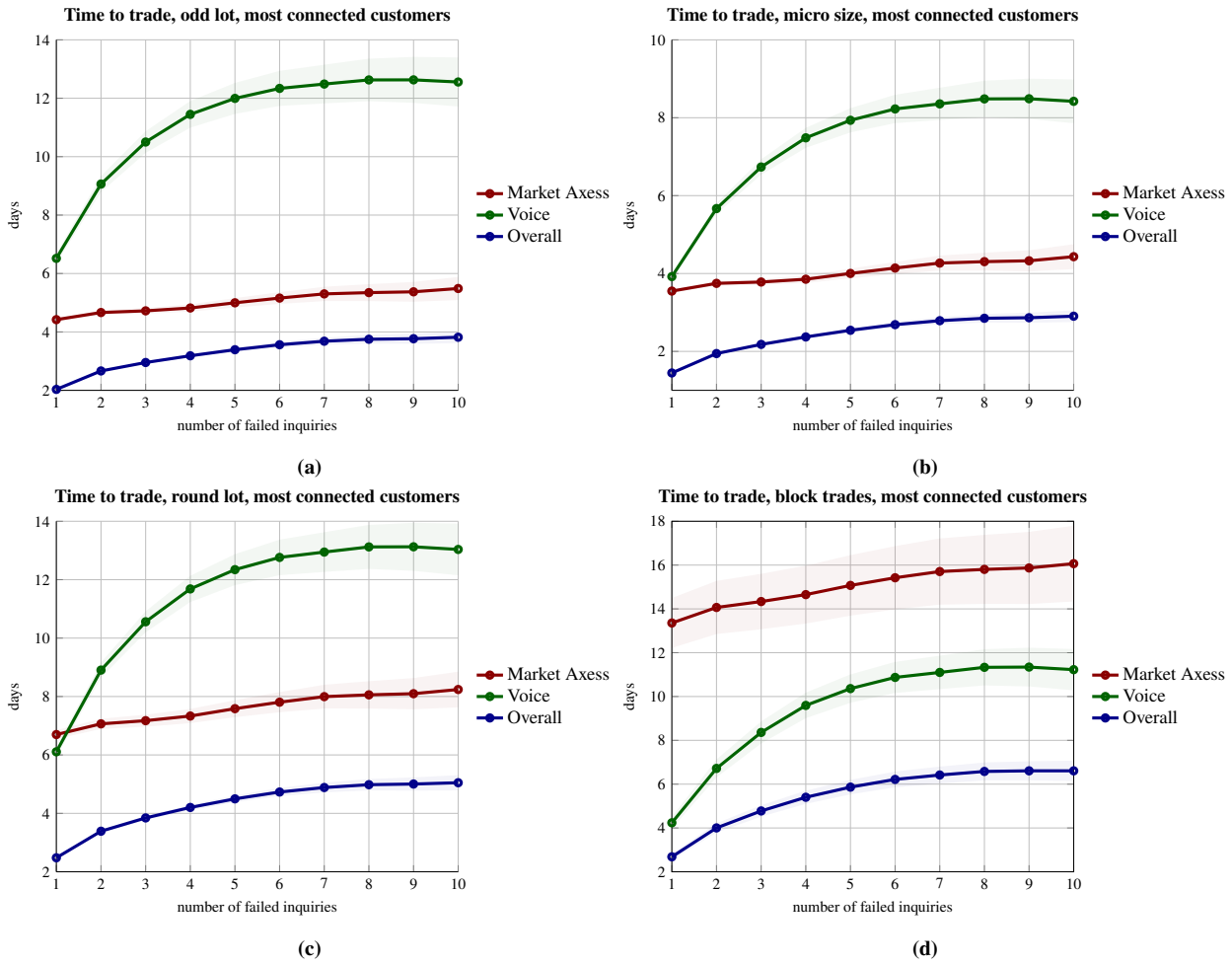


Figure 7. Estimated conditional time to trade from the MLE: MKTX vs. voice
This figure compares the estimated time to trade from Equation (5), conditional on the number of failed inquiries in MKTX vs. voice for the baseline (the top left panel), and different size categories. "Micro size" takes the value of 1 if the quantity of dealer response is below $100,000, and zero otherwise; "odd lot" takes the value of 1 if the quantity of dealer response is between $100,000 and $1 million, and zero otherwise; "round lot" takes the value of 1 if the quantity of dealer response is between $1 million and $5 million, and zero otherwise; "block trade" takes the value of 1 if the quantity of dealer response exceeds $5 million, and zero otherwise.

It shows that the trading probability goes down as the number of failed inquiries increases. Panel (b) shows the inquiry intensity with MKTX, calculated based on the MLE, $\lambda(\theta_1' x) + \lambda(\theta_2' x)$. Panel (c) plots the best markdown or spread. It reveals that the best markdown increases as the number of failed inquiries increases. Panel (d) shows the expected number of dealer responses, estimated by Poisson regression.[18] It shows that, as the number of failed inquiries increases, the expected number of dealer responses falls.

[18]The associated regression results are presented in column (1) of Table 10.

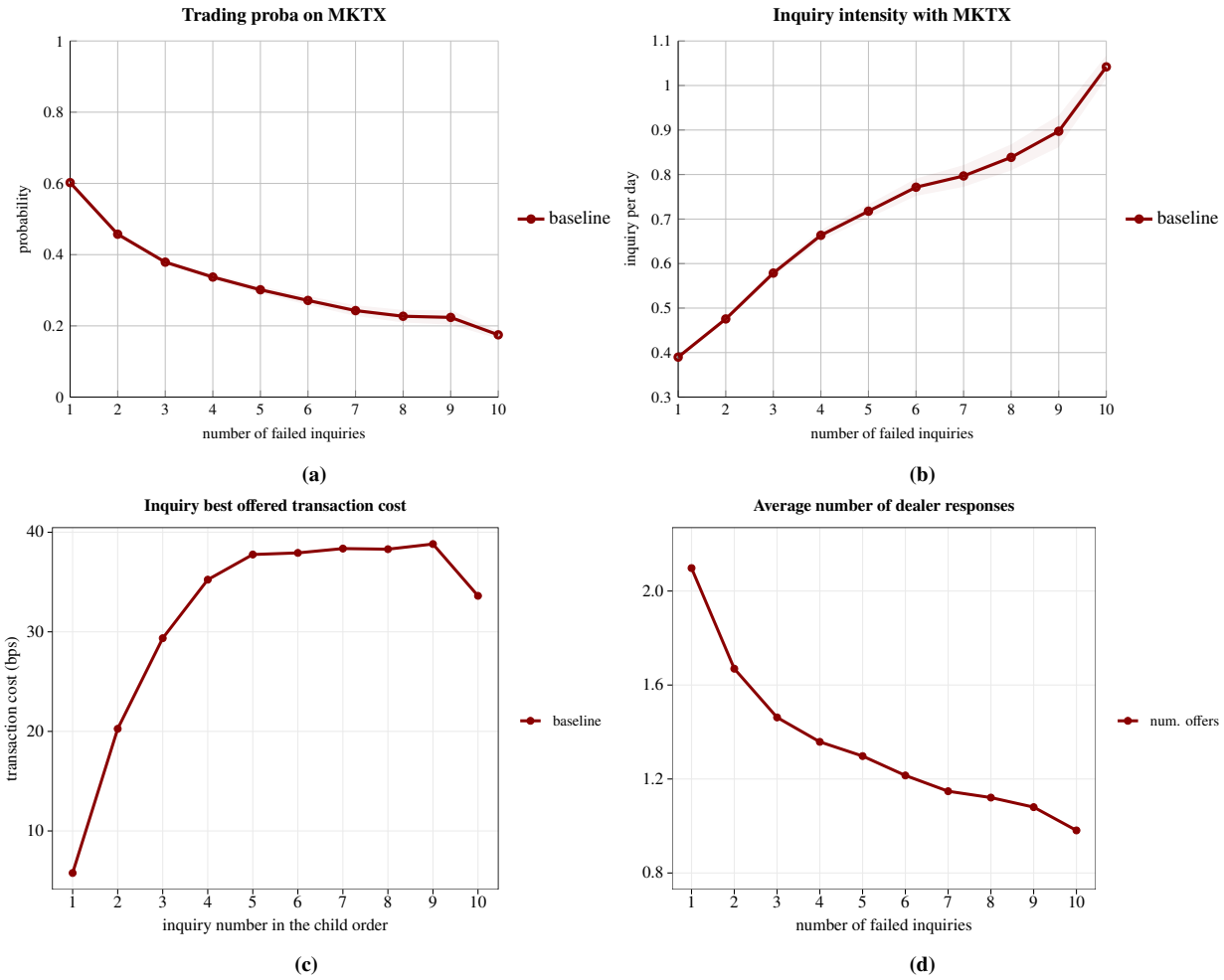


Figure 8. Unobserved characteristics
This figure plots, for the baseline category, the trading probability on MKTX in panel (a), the contact intensity with MKTX in panel (b), the markdown or best offered transaction cost for traded inquiries in panel (c), and the expected number of dealer responses in panel (d) as a function of the number of failed inquiries in child orders. The baseline category is an odd-lot purchase of an investment-grade bond, with high turnover, during normal times, for a connected investor, after one failed inquiry. As discussed in Section 2, we measure transaction cost as a markdown or markup relative to the benchmark provided by MKTX, called Composite+.

The dependence of outcome variables on the number of failed inquiries can be interpreted in two ways. The first explanation is that, as a customer fails more and more inquiries, the trading environment changes. Just to give one example, the increase in the spread for traded inquiries in panel (c) could be consistent with change in dealers' bidding behavior because of information leakage about this customer, as noted by Hendershott and Madhavan (2015), or because dealers learn that the customer was not able to elicit competitive offers, as in the "ringing-phone curse" described in Zhu (2011).

The second explanation is instead that child orders are heterogeneous in characteristics unobserved by the econometrician. Indeed, according to Lemma 1, the distribution of child orders along such unobserved characteristics changes systematically with the number of failed inquiries. For example, suppose that child orders differ in distress costs, c. Then, since child orders with higher distress costs have larger reservation markdown, $m^\star$, they trade with higher probability. This implies that, conditional on a larger number of failed inquiries, the sample becomes selected towards low distress cost child orders that trade with lower probability. Hence, heterogeneity in distress costs could explain why a child order time to trade appears to increase with the number of failed inquiries.

To tell these two hypotheses apart, we study the dependence of outcome variables on the number of failed inquiries in two ways: controlling for observed trade characteristics and controlling for child-order fixed effects. If unobserved child order characteristics explain the dependence of outcome variables on the number of failed inquiries, then the dependence should disappear after controlling for child order fixed effects. Indeed, when we control for child order fixed effects, we keep *all* child order characteristics fixed, whether they are observed or not.

[Table 10 about here.]

Table 10 shows the Poisson regression results when the outcome variable is the number of dealer responses. In column (1), we control for observed trade characteristics. We find that holding all observed trade characteristics constant, increasing the number of inquiries from 1 to 2, reduces the number of dealer responses by approximately 27% ($= 1 - e^{-0.311}$). Second, in column (2), we use child order fixed effects instead of trade characteristics. Now changing the number of inquiries has much more muted impact on the number of dealer responses: increasing the number of inquiries from 1 to 2, actually *increases* the number of dealer responses by 3.7% ($= 1 - e^{-0.0361}$). The results in Table 10 provide evidence in favor of the hypothesis that child orders differ in unobserved characteristics.

In Table 11, we do the same but for another variable: the spread (transaction cost) of traded inquiries. As discussed in Section 2, we measure transaction cost as a markdown or markup relative

to the CP+ benchmark provided by MKTX.[19] The evidence in column (1) is consistent with panel (b) of Figure 8, showing that spreads rise as the number of inquiries increases. However, when we control for potentially unobserved characteristics with child order fixed effects, in column (2), we obtain a very different picture: the spreads of traded inquiries are much more stable as the number of inquiries within a child order changes and, if anything, go slightly in the opposite direction.

[Table 11 about here.]

These regression results suggest that unobserved characteristics are a likely explanation of the dependence of outcome variables on the number of failed inquiries. But they do not shed light on the nature of these unobserved characteristics. However, our model can help make some indirect inference about these. For example, we can easily reject heterogeneity in distress cost: while this would explain why the trading probability decreases in the number of failed inquiries, as in panel (a) of Figure 8, it would be hard to reconcile with the observation that the inquiry intensity increases.

Which specific unobserved characteristics could be qualitatively consistent with the evidence in the four panels of Figure 8 and with the evidence on time to trade presented above? Panel (b) suggests that, as the number of failed inquiries increase, the sample becomes more and more selected towards child orders with high inquiry intensity on MKTX. According to Lemma 1, this is consistent with a McCall (1970) model in which investors are heterogeneous in their inquiry intensity on the electronic market, λ_e, because, in the child-order tree, the probability of a failed inquiry

$$\pi_1 = \frac{\lambda_e \left(\sum_j q^j \left[1 - G_e(m^\star) \right]^j \right)}{\lambda_e + \lambda_v G_v(m^\star) + \gamma},$$

is increasing in λ_e, for two reasons. First, the probability of an inquiry on the electronic market, $\lambda_e/(\lambda_e + \lambda_v G_v(m^\star) + \gamma)$, evidently increases with λ_e. Second, from Proposition 1, the reservation markdown, $m^\star$, decreases with λ_e, since customers who make more frequent inquiries have a larger

[19] As an alternative measure, we also compute the trading cost measure in Hendershott and Madhavan (2015), which uses the last inter-dealer trade as the reference price for a given bond instead of CP+. Results remain qualitatively similar using this alternative transaction cost measure.

option value of continuing their search. Panel (a) is consistent with this heterogeneity in λ_e too, since the trading probability on the electronic market declines with the number of failed inquiries.

However, Figure 8 is inconsistent with the hypothesis that child orders are *only* heterogeneous in their inquiry intensity λ_e. Indeed, panel (c) shows that the best offered transaction cost in a given inquiry on MKTX increases with the number of failed inquiries. This suggests that child orders are heterogeneous in another dimension: the distribution of their best offer on the electronic market. Panel (d) suggests a precise reason why this distribution may differ across child orders: the average number of dealer responses declines with the number of failed inquiries. Taken together, panels (a) through (d) of Figure 8 suggest that, after controlling for observable characteristics, child orders differ in two dimensions: first in their inquiry intensity, and second in the number of responses they elicit from dealers.

To formally establish that these two dimensions of heterogeneity are qualitatively consistent with the empirical observations in this paper, consider the McCall model where, for simplicity, child orders trade only on MKTX. Assume the distribution of dealer responses is Poisson with parameter μ. Suppose there are two types of child orders: type A with high inquiry intensity and low number of responses, and type B with low inquiry intensity and high number of responses. Let the associated parameters for these two types be $\lambda_A > \lambda_B$ and $\mu_A < \mu_B$. Denote the associated distribution of best offers by $F(m \mid \mu_A)$ and $F(m \mid \mu_B)$, where

$$F(m \mid \mu) = \sum_{j \geq 0} e^{-\mu} \frac{\mu^j}{j!} \left[1 - (1 - G(m))^j\right] = 1 - e^{-\mu G(m)}.$$

Finally, let the associated reservation markdowns be $m_A^\star$ and $m_B^\star$. Then we obtain the following Lemma.

Lemma 3 *Suppose that* $\lambda_A > \lambda_B > \lambda_A F(m_A^\star \mid \mu_A)$. *Then, as* $\mu_B \to \infty$*:*

$$\pi_{1A} > \pi_{1B}$$

$$F(m_A^\star \mid \mu_A) < F(m_B^\star \mid \mu_B)$$

$$\mathbb{E}_{\mu_A}[m] > \mathbb{E}_{\mu_B}[m] .$$

Taken together with Lemma 1, the first inequality in Lemma 3 implies that, conditional on a larger number of failed inquiries, the sample becomes more selected towards child orders with high inquiry intensity, but low number of dealer responses. This is consistent with both panel (b) and (d) of Figure 8. The second inequality in Lemma 3 makes it consistent with panel (a) of Figure 8, since child orders with high inquiry intensity but low number of dealer responses have lower trading probability. The third inequality makes it consistent with panel (c), since these child orders also trade at worse spread. Finally, the restriction that $\lambda_B > \lambda_A F(m_A^\star \mid \mu_A)$ ensures that these child orders have longer time to trade.

4 Conclusion

In this paper, we use data from a leading electronic trading platform to provide new and direct empirical evidence about search frictions in the OTC market for corporate bonds. We start from the observation that when a customer's inquiry on the platform fails to trade, the same customer often returns to the market shortly after to make subsequent inquiries for the same quantity of the same bond. We argue that the resulting sequence of repeated inquiries sheds light on the customers' sequential search process. We estimate that, after a failed inquiry, it takes customers between two and three days to trade. We show that this time to trade depends systematically on trade characteristics and trading venue (electronic vs. voice). We provide evidence consistent with unobserved characteristics being a likely reason for the dependence of outcome variables on the number of prior failed attempts to trade. Overall, our estimates can serve as useful inputs

into future quantitative applications of search models while also providing guidance for future theoretical explorations of the micro-foundations of search frictions in OTC markets.

References

Afonso, Gara, and Ricardo Lagos, 2015, Trade dynamics in the market for federal funds, *Econometrica* 83, 263–313.

Bao, Jack, Jun Pan, and Jiang Wang, 2011, The illiquidity of corporate bonds, *Journal of Finance* 66, 911–946.

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

Bessembinder, Hendrik, William Maxwell, and Kumar Venkataraman, 2006, Market transparency, liquidity externalities, and institutional trading costs in corporate bonds, *Journal of Financial Economics* 82, 251 – 288.

Bessembinder, Hendrik, Chester Spatt, and Kumar Venkataraman, 2020, A survey of the microstructure of fixed-income markets, *Journal of Financial and Quantitative Analysis* 55, 1–45.

Biais, Bruno, Johan Hombert, and Pierre-Olivier Weill, 2014, Equilibrium pricing and trading volume under preference uncertainty, *Review of Economic Studies* 81, 1401–1437.

Biais, Bruno, and Pierre-Olivier Weill, 2009, Liquidity shocks and order book dynamics, Working paper, TSE and UCLA.

Brancaccio, Giulia, and Karam Kang, 2021, Search frictions and product design in the municipal bond market, Working paper, NYU and CMU.

Czech, Robert, and Gábor Pintér, 2020, Informed trading and the dynamics of client-dealer connections in corporate bond markets, Working paper, Bank of England.

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

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

Edwards, Amy K., Lawrence E. Harris, and Michael S. Piwowar, 2007, Corporate bond market: Transaction costs and transparency, *Journal of Finance* 62, 1421–1451.

Flinn, Christopher, and James Heckman, 1982, New methods for analyzing structural models of labor force dynamics, *Journal of Econometrics* 18, 115–168.

Garbade, Kenneth D., and William L. Silber, 1976, Price dispersion in the government securities market, *Journal of Political Economy* 84, 721–740.

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

Goldstein, Michael A., Edith S. Hotchkiss, and Erik R. Sirri, 2007, Transparency and liquidity: A controlled experiment on corporate bonds, *Review of Financial Studies* 20, 235–273.

Hendershott, Terrence, Dan Li, Dmitry Livdan, and Norman Schürhoff, 2020, True cost of immediacy, Working paper, UC Berkeley, SFI, and FRB.

Hendershott, Terrence, Dmitry Livdan, and Norman Schürhoff, 2021, All-to-all liquidity in corporate bonds, Working paper, UC Berkeley and SFI.

Hendershott, Terrence, and Ananth Madhavan, 2015, Click or call? auction versus search in the over-the-counter market, *Journal of Finance* 70, 419–447.

Honoré, Bo H., and Adriana Lleras-Muney, 2006, Bounds in competing risks models and the war on cancer, *Econometrica* 74, 1675–1698.

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

Katz, Lawrence F., and Bruce D. Meyer, 1990, Unemployment insurance, recall expectations, and unemployment outcomes, *Quarterly Journal of Economics* 105, 973–1002.

Kiefer, Nicholas M, and George R Neumann, 1979, An empirical job-search model, with a test of the constant reservation-wage hypothesis, *Journal of Political Economy* 87, 89–107.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1336.

McCall, John J., 1970, Economics of information and job search, *Quarterly Journal of Economics* 84, 113–126.

O'Hara, Maureen, and Alex Zhou, 2021, The electronic evolution of corporate bond dealers, *Journal of Financial Economics* 140.

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

Schultz, Paul, 2001, Corporate bond trading costs: A peek behind the curtain, *Journal of Finance* 56, 677–698.

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

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

Wooldridge, Jeffrey M., 2010, *Econometric Analysis of Cross Section and Panel Data* (MIT Press).

Zhu, Haoxiang, 2011, Finding a good price in opaque over-the-counter markets, *Review of Financial Studies* 25, 1255–1285.

Appendix

A Omitted proofs

A.1 Proof of Lemma 1

Using the definition of H, we obtain that

$$dH(x \mid n) = \frac{\pi_1(x)\, d\mu(x \mid n-1)}{\int_{\underline{x}}^{\overline{x}} \pi_1(y)\, d\mu(y \mid n-1)} = \frac{\pi_1(x)\, dH(x \mid n-1)}{\int_{\underline{x}}^{\overline{x}} \pi_1(y)\, dH(y \mid n-1)}$$

where the first equality follows from the recursion $d\mu(x \mid n) = \pi_1(x)\, d\mu(x \mid n-1)$, and the second equality follows from dividing both the numerator and the denominator by $\int_{\underline{x}}^{\overline{x}} d\mu(x \mid n-1)$. Therefore:

$$\begin{aligned}\operatorname{sign}\left(H(x \mid n) - H(x \mid n-1)\right) &= \operatorname{sign}\left(\frac{\int_{\underline{x}}^{x} \pi_1(y)\, dH(y \mid n-1)}{\int_{\underline{x}}^{\overline{x}} \pi_1(y)\, dH(y \mid n-1)} - \int_{\underline{x}}^{x} dH(y \mid n-1)\right) \\ &= \operatorname{sign}\left(\int_{\underline{x}}^{x} \left[\pi_1(y) - \int_{\underline{x}}^{\overline{x}} \pi_1(z)\, dH(z \mid n-1)\right] dH(y \mid n-1)\right).\end{aligned}$$

Recall that $\pi_1(y)$ is strictly increasing. This implies that $\pi_1(y) - \int_{\underline{x}}^{\overline{x}} \pi_1(z)\, dH(z \mid n)$ is strictly increasing as well, strictly negative when $y = \underline{x}$, and strictly positive when $y = \overline{x}$. It follows that there is an x_0 such that $\pi_1(y) - \int_{\underline{x}}^{\overline{x}} \pi_1(z)\, dH(z \mid n-1) < 0$ for all $y < x_0$, and strictly positive for all $y > x_0$. Hence,

$$x \mapsto \int_0^x dH(y \mid n-1) \left[\pi_1(y) - \int_0^{\overline{x}} \pi_1(z)\, dH(z \mid n-1)\right]$$

is first decreasing and then increasing. Since this function is obviously equal to zero at the upper bound of its domain, $x = \overline{x}$, it follows that $H(x \mid n) \leq H(x \mid n-1)$, and we have established first-order stochastic dominance.

A.2 Proof of Lemma 3

Clearly, an increase in μ creates a first-order negative shift in $F(m \mid \mu)$ (i.e., markdown are lower). Hence the reservation markdown is decreasing in μ. Going to the limit in the reservation markdown equation (3), noting that $\lim_{\mu\to\infty} F(m \mid \mu) = 1$ for all $m > 0$, we obtain that

$$\lim_{\mu\to\infty} m^\star = \frac{c}{r+\gamma+\lambda} > 0.$$

But $F(m \mid \mu) \to 1$ for $m > 0$, i.e., the trading probability becomes arbitrarily close to 1, and the expected markdown becomes arbitrarily close to zero. The result follows.

Tables

Table 1. Responses of a traded and an untraded inquiry
Panel (a) provides dealers' disclosed responses for a traded inquiry submitted on 08/15/2017 to buy $300,000 of an 11-year, 3.824% investment-grade (USHG) bond issued on 01/17/2017 by Bank of America. The customer received 6 responses, all from dealers, whose anonymized IDs are provided in column (6). Response level (spread over Treasuries for USHG in MKTX) for each dealer response is reported in column (7). In column (10), the response status "Done" flags the response that the submitter accepted, the response status "Cover" flags the second best offer, and the response status "Missed" flags the rest of the responses that the submitter rejected. Panel (b) provides dealer disclosed responses for an untraded inquiry submitted on 08/17/2017 to buy $490,000 of the same bond in panel (a). The customer received 9 responses, all from dealers, whose anonymized IDs and response levels are reported in columns (6) and (7), respectively. The response status "DNT" for this inquiry in column (9) indicates that the inquiry did not trade.

Panel (a): Responses to a traded inquiry on 08/15/2017

Cust. ID (1)	Bond CUSIP (2)	Trade Side (3)	Submit Time (4)	Resp. ID (5)	Dealer ID (6)	Resp. Level (7)	Resp. Quant. (8)	Resp. Status (9)
127	06051GGF0	Buy	08:07:06	1	15420	126.37	300	Missed
127	06051GGF0	Buy	08:07:06	2	16323	129.70	300	Done
127	06051GGF0	Buy	08:07:06	3	11595	128.00	300	Missed
127	06051GGF0	Buy	08:07:06	4	16664	128.05	300	Missed
127	06051GGF0	Buy	08:07:06	5	10392	128.32	300	Missed
127	06051GGF0	Buy	08:07:06	6	12867	128.70	300	Cover

Panel (b): Responses to an untraded inquiry on 08/17/2017

Cust. ID (1)	Bond CUSIP (2)	Trade Side (3)	Submit Time (4)	Resp. ID (5)	Dealer ID (6)	Resp. Level (7)	Resp. Quant. (8)	Resp. Status (9)
127	06051GGF0	Buy	09:56:49	1	15420	125.32	490	DNT
127	06051GGF0	Buy	09:56:49	2	11122	125.70	490	DNT
127	06051GGF0	Buy	09:56:49	3	16377	124.70	490	DNT
127	06051GGF0	Buy	09:56:49	4	12867	125.70	490	DNT
127	06051GGF0	Buy	09:56:49	5	16323	126.20	490	DNT
127	06051GGF0	Buy	09:56:49	6	16664	125.31	490	DNT
127	06051GGF0	Buy	09:56:49	7	10392	125.32	490	DNT
127	06051GGF0	Buy	09:56:49	8	11684	127.01	490	DNT
127	06051GGF0	Buy	09:56:49	9	13910	126.71	490	DNT

Table 2. Cluster of inquiries
This table provides details on the inquiries composing the cluster for an 11-year, 3.824% investment-grade bond issued on 01/17/2017 by Bank of America over a six-month period in 2017, depicted in Figure 1.

Inquiry ID (1)	Cust. ID (2)	Bond CUSIP (3)	Trade Side (4)	Submit Time (5)	Requested Quantity (6)	Inquiry Traded? (7)	Parent Order # (8)	Child Order # (9)
1	127	06051GGF0	Buy	08/15/2017 08:07:06	300	Yes	1	1
2	127	06051GGF0	Buy	08/17/2017 09:56:49	490	No	1	2
3	127	06051GGF0	Buy	08/17/2017 13:57:19	490	Yes	1	2
4	127	06051GGF0	Buy	08/18/2017 08:35:20	290	No	1	3
5	127	06051GGF0	Buy	08/21/2017 08:45:43	290	Yes	1	3
6	127	06051GGF0	Buy	08/23/2017 11:11:38	680	Yes	1	4

Table 3. Trade probability on MKTX: inquiry vs. child order level
This table presents logit regression results of whether trade occurs as the dependent variable and indicators for trade and customer characteristics as independent variables, defined in Table 8. Column (1) presents the regression at the child order level and the corresponding inquiry level estimates are presented in column (2). We rank customers into deciles according to the number of dealer responses they receive, after controlling for inquiry size, fraction of requests for sell trades and HY bonds. "Connected decile 9" is an indicator for the customer being in decile 9, and similarly for other "Connected" indicators. Heteroskedasticity-robust standard-errors are reported in parentheses.

Dependent Variables:	child is traded	inq is traded
Model:	(1)	(2)
Variables		
(Intercept)	2.462***	2.019***
	(0.0029)	(0.0025)
Micro size	0.4815***	0.4187***
	(0.0021)	(0.0018)
Round lot	-0.3408***	-0.1795***
	(0.0035)	(0.0032)
Block trade	-0.6760***	-0.3941***
	(0.0109)	(0.0101)
Sell	0.6565***	0.6514***
	(0.0020)	(0.0017)
HY	-0.4866***	-0.4732***
	(0.0024)	(0.0021)
Covid	-0.8599***	-0.9448***
	(0.0053)	(0.0046)
Old age	-0.1863***	-0.1803***
	(0.0021)	(0.0018)
High time-to-maturity	-0.0659***	-0.0227***
	(0.0020)	(0.0017)
Low turnover	-0.3229***	-0.2827***
	(0.0027)	(0.0024)
Low amt outstanding	-0.5826***	-0.6091***
	(0.0020)	(0.0017)
Connected decile < 7	-2.296***	-2.186***
	(0.0028)	(0.0024)
Connected decile 7	-1.800***	-1.681***
	(0.0030)	(0.0027)
Connected decile 8	-1.204***	-1.159***
	(0.0030)	(0.0025)
Connected decile 9	-0.6484***	-0.7128***
	(0.0030)	(0.0024)
Fit statistics		
Observations	8,680,700	9,441,617
Squared Correlation	0.16503	0.17491
Pseudo R^2	0.16729	0.15811
BIC	6,969,783.3	8,986,339.2

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 4. Summary statistics

This table presents summary statistics for size, bond age and maturity, rating, and trade direction for all child orders (column 1), all inquiries (column 2), and child orders with at least one failed inquiry (column 3). "Sell" takes the value of 1 for a sale request, and zero otherwise; "HY" takes the value of 1 if the bond is high-yield, and zero otherwise; "Dealer-submitted" takes the value of 1 if the inquiry is submitted by a dealer, and zero otherwise.

	Child orders (all) (1)	Inquiries (all) (2)	Child orders (≥ 1 failed inq.) (3)
HY	0.17	0.18	0.26
Sell	0.52	0.51	0.42
Dealer-submitted	0.10	0.11	0.23
Size			
micro size (< $100k)	0.49	0.48	0.37
odd lot ($100k–1 million)	0.42	0.43	0.52
round lot ($1–5 million)	0.09	0.08	0.10
block trade (> $5 million)	0.01	0.01	0.01
Bond age distribution			
Average bond age	3.85	3.91	4.43
< 2 years	0.35	0.34	0.31
2–5 years	0.39	0.39	0.38
5–20 years	0.26	0.27	0.30
> 20 years	0.01	0.01	0.02
Bond maturity distribution			
Average maturity	12.43	12.53	13.71
< 2 years	0.002	0.002	0.002
2–5 years	0.07	0.07	0.06
5–20 years	0.73	0.73	0.68
> 20 years	0.20	0.20	0.25
Observations	9,861,143	11,020,815	2,774,478

Table 5. Child order event statistics

This table presents summary statistics about child order events. A child order can be viewed as a sequence of events, as depicted in Figure 2. Each element of the sequence is one of four possible events: an untraded inquiry on MKTX, a MKTX inquiry with trade, a voice trade, and, if the child order ends without a trade, an exit. By construction, the first event is always either an inquiry on MKTX, without or with trade. The first row shows the probability of a failed and successful inquiry on MKTX. The following rows provides the frequency distribution over the next event in the child order, conditional on the number of failed inquiries to date.

Event	Prob. MKTX inq. w/o trade (1)	Prob. MKTX inq. w trade (2)	Prob. voice trade (3)	Prob. exit (4)
First inquiry	0.28	0.72	N/A	N/A
After 1 failed inquiry	0.17	0.09	0.24	0.49
After 2 failed inquiries	0.34	0.10	0.17	0.39
After 3 failed inquiries	0.46	0.09	0.13	0.31
After 4 failed inquiries	0.55	0.08	0.10	0.26
After 5 failed inquiries	0.61	0.08	0.08	0.22
After 6 failed inquiries	0.66	0.07	0.07	0.20
After 7 failed inquiries	0.69	0.06	0.07	0.18
After 8 failed inquiries	0.73	0.06	0.06	0.16
After 9 failed inquiries	0.74	0.05	0.05	0.15
After 10 failed inquiries	0.77	0.05	0.05	0.14

Table 6. Child orders statistics: Inter-arrival times.
This table presents summary statistics about time between child order events (in business days). A child order can be viewed as a sequence of events, as depicted in Figure 2. Each element of the sequence is one of four possible events: an untraded inquiry on MKTX, a MKTX inquiry with trade, a voice trade, and, if the child order ends without a trade, an exit. Columns (1)–(3) present time, in business days, to an untraded inquiry on MKTX, a MKTX trade, and a trade on voice across child orders, conditional on the number of failed inquiries to date.

	Time to MKTX inq. w/o trade (1)	Time to MKTX inq. w trade (2)	Time to voice trade (3)
After 1 failed inquiry	0.82	0.65	1.04
After 2 failed inquiries	0.87	0.82	1.34
After 3 failed inquiries	0.85	0.85	1.46
After 4 failed inquiries	0.84	0.88	1.53
After 5 failed inquiries	0.82	0.85	1.56
After 6 failed inquiries	0.80	0.87	1.56
After 7 failed inquiries	0.78	0.89	1.63
After 8 failed inquiries	0.77	0.84	1.59
After 9 failed inquiries	0.75	0.86	1.44
After 10 failed inquires	0.72	0.88	1.44

Table 7. The unconditional Maximum Likelihood Estimator

This table presents estimation results for the unconditional MLE, where the only control is a constant for event $k \in \{1, \ldots, K\}$, with $K = 4$. Event $k = 1$ is an inquiry on MKTX without trade, $k = 2$ is an inquiry on MKTX with trade, $k = 3$ is a voice trade, and $k = 4$ is an exit. Robust standard errors as explained in Chapter 12.5.1 of Wooldridge (2010) are reported in parentheses. Our sample has $N = 2,383,637$ observations.

Event	MKTX inq. w/o trade (1)	MKTX inq. w trade (2)	voice trade (3)	exit (4)
	-3.59^{***}	-4.211^{***}	-3.40^{***}	-2.93^{***}
	(2.76×10^{-6})	(4.67×10^{-6})	(2.11×10^{-6})	(1.71×10^{-6})

Robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 8. The estimated coefficients of the MLE, part 1: trade characteristic dummies

This table presents the first part of our estimation results for the MLE, conditional on trade characteristics (this table) and the number of failed inquiries in the child order to date (in Table 9). "Sell" takes the value of 1 for a sale request, and zero otherwise; "Ba1 to Caa3" takes the value of 1 if the bond's Moody's rating is between Ba1 and Caa3; "Ca to C" is similarly defined; "COVID" takes the value of 1 if the RFQ is submitted in March 2020, and zero otherwise; "Old" takes the value of 1 if the bond's age is above the 75th percentile of the distribution, and zero otherwise; "Turnover below median" takes the value of 1 if the bond's quarterly turnover is below median, and zero otherwise; "High time-to-maturity" takes the value of 1 if the bond's time to maturity is above the sample median, and zero otherwise; "Low amt outstanding" takes the value of 1 if the bond's amount outstanding is below the sample median, and zero otherwise; "Micro size" takes the value of 1 if the quantity of dealer response is below \$100,000, and zero otherwise; "Odd lot" takes the value of 1 if the quantity of dealer response is between \$100,000 and \$1 million, and zero otherwise; "Round lot" takes the value of 1 if the quantity of dealer response is between \$1 million and \$5 million, and zero otherwise; "Block trade" takes the value of 1 if the quantity of dealer response exceeds \$5 million, and zero otherwise. We rank customers into deciles according to the number of dealer responses they receive, after controlling for inquiry size, fraction of requests for sell trades and HY bonds. "Connected decile 9" is an indicator for the customer being in decile 9, and similarly for other "Connected" indicators. Robust standard errors as explained in Chapter 12.5.1 of Wooldridge (2010) are reported in parentheses. Our sample has $N = 2,383,637$ observations.

Event	MKTX inq. w/o trade (1)	MKTX inq. w trade (2)	voice trade (3)	exit (4)
(Intercept)	−4.06***	−3.65***	−3.38***	−2.91***
	(0.0056)	(0.0065)	(0.0049)	(0.0044)
Sell	0.121***	0.632***	0.354***	−0.029***
	(0.0036)	(0.0048)	(0.0033)	(0.003)
Ba1 to Caa3	−0.00544*	−0.0246***	0.0769***	−0.203***
	(0.0042)	(0.0056)	(0.0039)	(0.0035)
Ca to C	0.00523	−0.4***	0.259***	−0.255***
	(0.041)	(0.065)	(0.036)	(0.036)
COVID	−0.0705***	−0.483***	−0.412***	−0.167***
	(0.0074)	(0.0098)	(0.0065)	(0.0059)
Old	0.0157***	−0.108***	−0.0549***	0.045***
	(0.0036)	(0.0047)	(0.0032)	(0.0029)
Turnover below median	0.00936**	−0.106***	−0.0617***	0.164***
	(0.0041)	(0.0056)	(0.0041)	(0.0034)
High time-to-maturity	−0.0111***	0.0539***	−0.0852***	0.0804***
	(0.0036)	(0.0048)	(0.0033)	(0.0029)
Low amt outstanding	0.155***	−0.264***	−0.294***	0.179***
	(0.0036)	(0.0047)	(0.0032)	(0.0029)
Micro size	0.0277***	0.213***	0.399***	−0.311***
	(0.0035)	(0.0047)	(0.0034)	(0.003)
Round lot	−0.186***	−0.407***	−0.0374***	0.366***
	(0.0075)	(0.0099)	(0.0071)	(0.0056)
Block trade	−0.436***	−1.07***	0.112***	0.594***
	(0.03)	(0.04)	(0.026)	(0.021)
Connected decile < 7	0.0436***	−1.63***	−0.219***	0.0469***
	(0.0049)	(0.0072)	(0.0044)	(0.0038)
Connected decile 7	0.0322***	−1.14***	−0.0113**	0.0297***
	(0.0059)	(0.0078)	(0.0052)	(0.0046)
Connected decile 8	0.286***	−0.503***	0.611***	0.0966***
	(0.006)	(0.0076)	(0.0057)	(0.0051)
Connected decile 9	0.113***	−0.318***	−0.135***	−0.123***
	(0.0052)	(0.006)	(0.0049)	(0.0043)
Failed inquiry controls	Yes	Yes	Yes	Yes

Robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 9. The estimated coefficient of the MLE, part 2: the failed inquiries dummies
This table presents the second part of our estimation results for the MLE, conditional on trade characteristics (in Table 8) and the number of failed inquiries in the child order to date (this table). Event $k = 1$ is an inquiry on MKTX without trade, $k = 2$ is an inquiry on MKTX with trade, $k = 3$ is a voice trade, and $k = 4$ is an exit. "Failed j" takes the value of 1 if the number of failed inquiries in the child order to date is equal to j, and zero otherwise. Robust standard errors as explained in Chapter 12.5.1 of Wooldridge (2010) are reported in parentheses. Our sample has $N = 2,383,637$ observations.

Event	MKTX inq. w/o trade (1)	MKTX inq. w trade (2)	voice trade (3)	exit (4)
(Intercept)	-4.06^{***}	-3.65^{***}	-3.38^{***}	-2.91^{***}
	(0.0056)	(0.0065)	(0.0049)	(0.0044)
Failed 2	0.509^{***}	-0.0764^{***}	-0.497^{***}	-0.345^{***}
	(0.0045)	(0.0063)	(0.0045)	(0.0039)
Failed 3	0.841^{***}	-0.0685^{***}	-0.729^{***}	-0.567^{***}
	(0.0062)	(0.01)	(0.0081)	(0.0068)
Failed 4	1.04^{***}	-0.0485^{***}	-0.943^{***}	-0.738^{***}
	(0.0081)	(0.015)	(0.013)	(0.011)
Failed 5	1.17^{***}	-0.082^{***}	-1.08^{***}	-0.858^{***}
	(0.01)	(0.022)	(0.02)	(0.015)
Failed 6	1.29^{***}	-0.114^{***}	-1.22^{***}	-0.949^{***}
	(0.012)	(0.029)	(0.028)	(0.021)
Failed 7	1.36^{***}	-0.193^{***}	-1.23^{***}	-1.1^{***}
	(0.015)	(0.038)	(0.035)	(0.028)
Failed 8	1.43^{***}	-0.21^{***}	-1.35^{***}	-1.21^{***}
	(0.017)	(0.048)	(0.047)	(0.036)
Failed 9	1.5^{***}	-0.156^{***}	-1.42^{***}	-1.24^{***}
	(0.019)	(0.056)	(0.06)	(0.044)
Failed ≥ 10	1.71^{***}	-0.253^{***}	-1.34^{***}	-1.3^{***}
	(0.011)	(0.036)	(0.033)	(0.027)
Trade char. controls	Yes	Yes	Yes	Yes

Robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 10. Poisson model for the number of dealer responses
This table presents Poisson regression estimates for number of dealer responses on indicators for the inquiry number in child orders. "Inquiry j" takes the value of 1 if it is the jth inquiry in the child order. In column (1) we include trade characteristics described in Table 8. In column (2), we control for the unobserved child order characteristics by adding child order fixed effects to the regression. The sample excludes inquiries submitted by dealers.

Dependent Variable:	number of dealer responses	
Model:	(1)	(2)
Variables		
(Intercept)	1.903***	
	(0.0003)	
Inquiry 2	-0.3110***	0.0361***
	(0.0008)	(0.0006)
Inquiry 3	-0.4241***	0.0607***
	(0.0016)	(0.0012)
Inquiry 4	-0.4724***	0.0670***
	(0.0027)	(0.0019)
Inquiry 5	-0.4799***	0.0812***
	(0.0038)	(0.0026)
Inquiry 6	-0.4990***	0.0867***
	(0.0051)	(0.0034)
Inquiry 7	-0.5211***	0.0837***
	(0.0066)	(0.0044)
Inquiry 8	-0.5163***	0.1017***
	(0.0081)	(0.0053)
Inquiry 9	-0.5156***	0.1065***
	(0.0097)	(0.0064)
Inquiry $\geq$ 10	-0.4973***	0.1055***
	(0.0055)	(0.0069)
Trade char. controls	Yes	
Fixed-effects		
child order		Yes
Fit statistics		
Observations	9,455,325	9,108,063
Squared Correlation	0.33738	0.99172
Pseudo R^2	0.14526	0.36693
BIC	45,117,005.9	165,520,283.0
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*		

Table 11. Unobserved heterogeneity: Transaction costs
This table reports the estimates of regressing inquiry spreads on indicators for inquiries in child orders. In column (1), we include trade characteristics described in Table 8 and year-month fixed effects. In column (2), in addition to indicators for inquiry number in child orders, we add year-month and child order fixed effects to control for unobserved heterogeneity. As discussed in Section 2, we measure transaction cost as a markdown or markup relative to the benchmark provided by MKTX, called Composite+.

Dependent Variable:	inq. best offered transaction cost (bps)	
Model:	(1)	(2)
Variables		
Inquiry 2	8.008***	-2.307***
	(0.4365)	(0.0600)
Inquiry 3	12.12***	-2.254***
	(0.8524)	(0.1282)
Inquiry 4	13.26***	-2.154***
	(1.536)	(0.2092)
Inquiry 5	14.18***	-1.311***
	(2.415)	(0.2935)
Inquiry 6	12.81***	-2.073***
	(3.402)	(0.3876)
Inquiry 7	11.23**	-1.924***
	(4.410)	(0.5107)
Inquiry 8	11.32**	-0.6653
	(5.146)	(0.5839)
Inquiry 9	11.47**	-0.1492
	(5.077)	(0.6896)
Inquiry $\geq$ 10	7.668	-1.017
	(4.992)	(0.8269)
Trade char. controls	Yes	
Fixed-effects		
child order		Yes
year-month	Yes	Yes
Fit statistics		
Observations	8,212,803	8,218,662
R^2	0.20256	0.95339
Within R^2	0.18961	0.04535

Clustered standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Algorithmic Pricing and Liquidity in Securities Markets*

Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo

March 10, 2023

Abstract

We let "Algorithmic Market-Makers" (AMs), using Q-learning algorithms, choose prices for a risky asset when their clients are privately informed about the asset payoff. We find that AMs learn to cope with adverse selection and to update their prices after observing trades, as predicted by economic theory. However, in contrast to theory, AMs charge a mark-up over the competitive price, which declines with the number of AMs. Interestingly, markups tend to decrease with AMs' exposure to adverse selection. Accordingly, the sensitivity of quotes to trades is stronger than that predicted by theory and AMs' quotes become less competitive over time as asymmetric information declines.

Keywords: Algorithmic pricing, Market Making, Adverse Selection, Market Power, Reinforcement learning.

JEL classification: D43, G10, G14.

*Correspondence: colliard@hec.fr, foucault@hec.fr, lovo@hec.fr. All authors are at HEC Paris, Department of Finance, 1 rue de la Libération, 78351 Jouy-en-Josas, France. We are grateful to participants in "The Microstructure Exchange", the Microstructure Asia Pacific Online Seminar, and seminars at the University of Copenhagen, University Paris 1, and HEC Paris for helpful comments and suggestions. We thank Olena Bogdan, Amine Chiboub, Chhavi Rastogi and Andrea Ricciardi for excellent research assistance. This work was supported by the French National Research Agency (F-STAR ANR-17-CE26-0007-01, ANR EFAR AAP Tremplin-ERC (7) 2019), the Investissements d'Avenir Labex (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047), the Chair ACPR/Risk Foundation "Regulation and Systemic Risk", the Natixis Chair "Business Analytics for Future Banking".

Introduction

Firms (e.g., retailers, airlines, hotels, energy providers etc.) increasingly rely on algorithms to set the price of their products.[1] This evolution reflects efficiency and predictive gains of artificial intelligence but it generates new concerns, in particular about price discrimination and tacit collusion among algorithms (see MacKay and Weinstein (2022), CMA (2018), OECD (2017)).[2] Surprisingly, this worry has not been expressed for market makers in securities markets even though proprietary trading firms (market makers such as Citadel, Virtu, Jane Street, etc.) began using pricing algorithms at least two decades ago and now dominate liquidity provision in exchanges.[3]

Is this lack of concern justified? Is tacit collusion among pricing algorithms more difficult in securities markets? To study this question, we consider a framework in which "algorithmic market makers" (AMs) compete in prices and are at risk of trading with better informed investors. Our setting is, by design, very similar to standard models of market making with asymmetric information (in the spirit Glosten and Milgrom (1985) and Kyle (1985)). However, in contrast to these models, we assume that quotes are posted by AMs that set their quotes using Q-learning algorithms, a special type of reinforcement learning algorithm (often mentioned as the type of algorithms used for pricing decisions; see CMA (2018)). We focus on whether Q-learning algorithms cope with adverse selection, learn to account for the information contained in trades in choosing prices and whether their prices are competitive. To our knowledge, our paper is the first to analyze how Q-learning algorithms behave in the presence of asymmetric information (an important feature of trading in securities markets).

In our framework, AMs simultaneously post offers in response to clients' requests to buy one share of a risky asset. Clients' valuation for the asset is the sum of the payoff of the asset (a common value component) and a component specific to each client (a private valuation component). Clients

[1] For instance, Chen *et al.* (2016) find that more than 500 Amazon third-party sellers on Amazon marketplace were using algorithms to price their products.

[2] For instance, MacKay and Weinstein (2022) write: *"The explosion in the use of pricing algorithms over the past decade has sparked concerns about the effect on competition and consumers [...]."*.

[3] These firms are often referred to as "high-frequency market makers" because their algorithms (and hardware equipments) generate very frequent new orders (quotes, cancellations etc.). Menkveld (2013) finds that, in 2007-2008, a single high-frequency market maker accounts for about 15% of total trading volume in Dutch stocks (and more than 60% on one of the trading platforms for these stocks). Brogaard *et al.* (2015) find that fast traders on the Stockholm Stock Exchange are primarily market makers, who account for 83% of all limit orders on this exchange (and 44% of trading volume).

arrive sequentially and each one trades with the dealers posting the best offer in response to her request, provided that this offer is less than her valuation.[4] As clients' demand for the asset increases with the common value, dealers are exposed to adverse selection (they are more likely to sell the asset when its payoff is high than when it is low). In the baseline case, a new realization of the asset payoff is drawn after each client's arrival.

AMs behave as follows. In each trading round, each AM starts with an assessment of the expected profit associated with each possible price, picks a price (on a grid) based on this assessment, and updates its assessment of the expected profit associated with this price by taking a weighted average (with pre-specified weights) of its realized profit at the end of the trading round and its prior assessment of its expected profit. The assessment of the expected profit associated with other possible prices is unchanged. In each round t, the AM picks the price that generates the largest expected profit according to its assessment with a given probability of "exploitation". Otherwise, it "explores" by picking at random (with equal probability for each possible price) another price. Exploration enables the AM to receive feedback about the profit generated by a price and therefore to "learn" the expected profit associated with this price.

This iterative process is repeated over a large number of "episodes" (each made of one trading round), which collectively constitute one "experiment". In a given experiment, the set of parameters (e.g., the number of AMs, the distribution of the asset payoff, and the distribution of each client's private valuations) is constant across episodes and forms the "environment". For each environment considered in our analysis (i.e., for a fixed set of parameters), we run $10,000$ experiments, each made of $200,000$ episodes. In each experiment we record each AM's quote, the transaction price, and the trading volume (0, or 1) in each episode.

In early episodes of a given experiment, AMs "learn" the expected profit associated with each possible price, which leads to significant volatility in prices. After a large number of episodes, their pricing strategy eventually "converges" in most experiments, in the sense that AMs keep playing the same price over a large number of episodes. However, this "long run" price can vary

[4] In electronic securities markets (e.g., electronic limit order books markets used in most of the major stock markets in the world), market makers compete in prices ("à la Bertrand") with no room for product differentiation. Price priority is strictly enforced, which guarantees that clients' orders are filled at the best price, as assumed in our analysis.

from one experiment to another. Thus, our analysis focuses on the *empirical distribution* of final outcomes (e.g., transaction prices and dealers' profits) across experiments (holding the environment constant). We study how this distribution varies with parameters of the environment (in particular the intensity of adverse selection) and we systematically compare final outcomes to those predicted by economic theory. When there are multiple dealers, the outcomes predicted by theory (e.g., transaction prices and profits) are those corresponding to the Bertrand-Nash equilibrium of the environment considered in our simulations (accounting for the fact that market makers must post their quotes on a grid).[5] When there is a single dealer, predicted outcomes are those corresponding to the equilibrium of the environment in which the dealer behaves as a monopolist.

We observe several interesting regularities. First, when there is a single AM, it does not necessarily learn the monopoly price and its average quote is smaller than the monopoly price. In contrast, when there are two AMs, they charge a price above the competitive price on average (even the smallest price in the distribution of observed prices is well above the competitive price). This markup decreases with the number of AMs and becomes close to zero only with 10 AMs.

Second, in all environments, AMs learn how *not* to be adversely selected. That is, they charge prices that (more than) cover adverse selection costs. However, when their exposure to adverse selection increases (either because the volatility of the asset payoff increases or the variance of clients' private valuations decreases), AMs tend to choose prices that are *more* competitive (in particular, their realized bid-ask spreads are smaller on average). This is particularly striking when the variance of clients' private valuations increases. In this case, AMs' offers (and therefore transaction prices) increase, even though the competitive (Nash-Bertrand) price decreases because adverse selection costs decline. Overall these findings suggest that adverse selection interacts with the way Q-learning algorithms learn in non-intuitive ways.

In existing models (Kyle (1985) or Glosten and Milgrom (1985)), market makers are assumed to learn the asset payoff from the trading history (the "order flow") in a Bayesian way. For this reason, holding the asset payoff constant, these models predict that dealers eventually discover asset payoffs. For instance, dealers' pricing errors (the average squared difference between the asset payoff

[5] In particular, as is well-known, price discreteness can generate multiple Bertrand-Nash equilibria. We take this into account in our analysis.

and the transaction price) decrease over time (trades) on average (see, for instance, Glosten and Milgrom (1985)). To study whether AMs can also discover asset values, we extend our baseline setting to allow for two trading rounds per episode. We find that AMs behave qualitatively like a Bayesian learner would. That is, after a buy (no trade), they increase (reduce) their offer in the second trading round. Thus, price discovery takes place, even though AMs have no knowledge of the data generating process and their learning process is not Bayesian. However, observing the outcome of the first period brings new information to the algorithms, which then face less adverse selection in the second period. As adverse selection curbs algorithms' rent-seeking behavior, prices become less competitive on average in the second period. Moreover, this effect is stronger after observing a trade than after observing no trade in the first period. In this sense, AMs seem to overreact to trades relative to a Bayesian learner.

In summary, our findings have two main implications. First, algorithmic market makers settle on non competitive prices, even though they operate in an environment where economic theory predicts competitive outcomes. This echoes findings in Hendershott *et al.* (2011) and Brogaard and Garriott (2019). The former find empirically that algorithmic trading (AT) *increases* dealers' expected profits net of adverse selection costs (realized bid-ask spreads). Commenting on this result, they write (on p.4): *"This is surprising because we initially expected that if AT improved liquidity, the mechanism would be competition between liquidity providers."* Brogaard and Garriott (2019) study the effect of high frequency market makers' entry on one trading platform for Canadian stocks. They find that this entry triggers a decrease in bid-ask spreads. However, the entry of two competitors is not sufficient to obtain the competitive outcome, in contrast to what standard models of market making predicts. This pattern is exactly what we find when we compare average bid-ask spreads across environments that only differ by the number of AMs. Our second main implication is that adverse selection induces AMs to post quotes that are *more* competitive. In a cross-section of assets, this means that *realized bid-ask spreads* for AMs (a measure of dealers' expected profits net of adverse selection costs) should be smaller in assets that are more exposed to informed trading. For instance, they should be smaller for stocks than for Treasuries (high frequency market makers are active in both types of assets), even though adverse selection costs are larger for stocks. This

implication also holds dynamically: as adverse selection is resolved over time we expect AMs to quote less competitively, contrary to the predictions of standard asymmetric information models.

It is worth stressing that we do not claim that market-making algorithms necessarily behave as our AMs.[6] This does not mean however that the patterns uncovered by our analysis are unlikely to hold in reality.[7] Our approach is to make stylized assumptions on pricing algorithms to develop predictions about their effects on securities markets. In particular, our assumptions capture that (some) algorithms used in practice rely on experimentation ("trial and error") but eventually experiment less and less, as experimentation is costly. We believe these properties are reasonable in a financial context. As explained above, this approach delivers predictions that are quite different from those of standard economic models in the same environment. To decide which approach has more explanatory power, the next step (which is beyond the scope of this paper) would be to test these predictions empirically.

The rest of the paper proceeds as follows. In the next section, we position our contribution in the literature. Section 2, we present the economic environment analyzed in our paper. In Section 3, we study the case in which each episode has a single trading round. In this case, our analysis focuses on how AMs' behavior differ from two benchmarks: (a) competitive behavior (Nash-Bertrand equilibrium) and (b) monopolistic behavior (monopoly prices). In Section 4, we study whether AMs can discover asset fair values by considering the case in which each episode has two trading rounds. Section 5 concludes.

1 Contribution to the literature

Our paper is related to the emerging literature on algorithmic pricing and the possibility for algorithms to sustain non competitive outcomes. Calvano *et al.* (2020) show that Q-learning algorithms

[6] There is not much guidance on the actual design of market making algorithms in reality because market making firms do not disclose information on their algorithms. Securities trading firms strongly push back a regulator's attempt to require disclosure of their computer codes for surveillance purpose. See *"US regulator declares 'dead' moves to seize HFT code"*, Financial Times, October 14, 2017. In the EU, proprietary trading firms must make sure that they take steps to insure that their algorithms will not lead to disordely markets. However, they do not have to disclose their algorithms to regulators.

[7] The behavior of market makers in existing economic models is also highly stylised and, in contrast to our approach here, they are assumed to have a complete knowledge of their environment (e.g., the distribution of asset payoffs, traders' valuations etc.). Yet, these models have explanatory power for the behavior of security prices at high frequency (see, for instance, Glosten and Harris (1988) and the subsequent literature using price impact regressions).

can learn dynamic collusive strategies in a repeated differentiated Bertrand game. Asker *et al.* (2021) and Abada *et al.* (2022) show that supra competitive prices can be reached in this type of environment even if dynamic strategies are ruled out, through what Abada *et al.* (2022) call "collusion by mistake".[8] Cartea *et al.* (2022a) and Cartea *et al.* (2022b) study different families of reinforcement learning algorithms and develop new methods to study which ones may lead to non Nash behavior in a market-making environment.[9] Banchio and Skrzypacz (2022) find that Q-learning algorithms post less competitive bids in first price auctions than in second price auctions. In contrast to our setting, bidders and sellers have a fixed valuation for the auctioned good and bidders are not exposed to adverse selection in their setting (they consider private value auctions). In sum, in line with other papers, we find that pricing algorithms relying on Q-learning can lead to non competitive outcomes even when dynamic strategies are ruled out and when price setters compete in prices. However, new to the literature, we find that adverse selection tends to mitigate this issue.[10] Moreover, to our knowledge, we are the first to study price discovery with Q-learning algorithms (an issue specific to securities markets).

Our paper also contributes to the literature on algorithmic trading in securities markets. The theoretical literature on this issue (e.g., Biais *et al.* (2015), Budish *et al.* (2015), Menkveld and Zoican (2017), Baldauf and Mollner (2020), etc.) has mainly focused on how the increase in the speed with which algorithms can respond to information increases or reduces liquidity suppliers' exposure to adverse selection, using traditional workhorses models (Glosten and Milgrom (1985) or Kyle (1985)). Yet, O'Hara (2015) calls for the development of new methodologies to study the effects of algorithms in financial markets, writing that as a result of algorithmic trading: *"the data that emerge from the trading process are consequently altered [...] For microstructure researchers, I believe these changes call for a new research agenda, one that recognizes how the learning models used in the past are lacking [...]."* Our paper responds to this call. Instead of modeling algorithmic traders as Bayesian learners, with an omniscient knowledge of the environment in which they operate, we

[8] This idea is in line with an earlier literature in machine learning showing that games between Q-learning algorithms do not necessarily converge to a Nash equilibrium (Wunder *et al.*, 2010).

[9] In particular, Cartea *et al.* (2022b) show that using a finer pricing grid (a lower "tick size") reduces the scope for collusion.

[10] Another rather unique feature of our setting is that, in our setting, the demand faced by pricing algorithms is stochastic. See also Hansen *et al.* (2021) and Cartea *et al.* (2022b) other settings in which selling algorithms face a stochastic demand elasticity, but without adverse selection.

model them as Q-learning algorithms. These algorithms learn by trial and error with almost no prior knowledge of the environment, which represents the polar opposite of standard Bayesian learning. Moreover, Q-learning is relatively simple and transparent, which makes it a good candidate for a workhorse model of algorithmic interaction, much like the Glosten-Milgrom environment is a workhorse model of market-making. This approach generates strikingly different implications for those of canonical Bayesian-learning models. In particular, price competition does not guarantee a competitive outcome and, maybe even more surprisingly, increased adverse selection can reduce dealers' rents.

2 The economic environment

In this section, we provide a general description of the economic environment considered in our experiments (Section 3.3). We consider the market for one risky asset with $t = 1, 2, ..., T$ episodes (one can think of them as "trading days"). Quotes in this market are posted by N dealers who trade with clients. Each episode has $\bar{\tau}$ trading rounds and the asset payoff $\tilde{v}$ is realized at the end of the last trading round in a given episode. This payoff has a binary distribution, $\tilde{v} \in \{v_L, v_H\}$, with $v_L \leq v_H$ and $\mu := \Pr(\tilde{v} = v_H) = \frac{1}{2}$. We denote $\Delta v = v_H - v_L$. Realizations of the asset payoffs are independent across episodes. In the rest of this section, we describe traders' actions and realized profits in a given episode.

In each trading round τ, a new trader (the "client") arrives to buy one share of the asset. The buyer's valuation for the asset is $v_\tau^C = \tilde{v} + \tilde{L}_\tau$, where $\tilde{L}_\tau$ is i.i.d across trading rounds. Clients' private valuations are assumed to be normally distributed with mean zero and variance σ^2. The buyer observes her valuation for the asset and requests quotes from the dealers, who simultaneously respond by posting their offers $a(\tau) = \{a_{1\tau}, ..., a_{N\tau}\}$. The ask price $a_{n\tau}$ is the price at which dealer n is ready to sell at most one share in trading round τ. We denote by (i) $a_\tau^{min} = \min_n\{a_{n\tau}\}$ the smallest ask price, (ii) $\mathcal{D}_\tau$ the set of dealers posting this price and (iii) z_τ the number of dealers in $\mathcal{D}_\tau$. The client buys the asset if the best offer is less than her valuation for the asset ($a_\tau^{min} \leq v_\tau^C$) and, in this case, she splits her demand among the z_τ dealers posting this price. Otherwise she does not trade.

Let denote by $V(a(\tau), \tilde{L}_\tau, \tilde{v})$ the volume of trade in round τ. It equals 1 if the client buys the asset, i.e., if the client valuation $\tilde{v} + \tilde{L}_\tau$ is not smaller than the lowest price a_τ^{min}, and it is zero otherwise. Let denote by $Z(a_{n\tau}, a(\tau))$ the fraction of the τ^{th} round trade executed by dealer n, that is, $Z(a_{n\tau}, a(\tau)) = \frac{1}{z_\tau}$ if $a_{n\tau} = a_\tau^{min}$ and is zero otherwise. In trading round τ, dealer n's realized trading volume is:

$$I(a_{n,\tau}, a(\tau), \tilde{L}_\tau, \tilde{v}) := V(a(\tau), \tilde{L}_\tau, \tilde{v}) Z(a_{n\tau}, a(\tau)), \tag{1}$$

and his realized profit is:

$$\Pi(a_{n\tau}, a(\tau), \tilde{L}_\tau, \tilde{v}) := I(a_{n,\tau}, a(\tau), \tilde{L}_\tau, \tilde{v})(a_\tau^{min} - \tilde{v}), \tag{2}$$

Hence, dealer n's total realized profit in a given episode is:

$$\sum_{\tau=1}^{\overline{\tau}} \Pi(a_{n\tau}, a(\tau), \tilde{L}_\tau, \tilde{v}). \tag{3}$$

In our setting, holding prices constant, dealers are more likely to sell the asset when the asset payoff is high than when the asset payoff is low. Indeed, conditional on a realization of v, the likelihood that a trade occurs in trading round τ is:

$$D(a_\tau^{min}, v) := Pr(a_\tau^{min} \leq v + \tilde{L}_\tau) = 1 - G(a_\tau^{min} - v), \tag{4}$$

which increases with v as $D(a_\tau^{min}, v_H) > D(a_\tau^{min}, v_L)$. Thus, dealers are exposed to adverse selection: they are more likely to sell the asset when its payoff is high than when it is low.

Finally, we denote by $\overline{\Pi}(a, \mu_\tau) = N\mathbb{E}[\Pi(a, a, \tilde{L}, \tilde{v})]$ the dealers' expected aggregate profit when they all post the same price a, and attach probability μ_τ to the event $\tilde{v} = v_H$. This gives

$$\bar{\Pi}(a, \mu_\tau) := \mu_\tau D(a, v_H)(a - v_H) + (1 - \mu_\tau) D(a, v_L)(a - v_L) \tag{5}$$

In the rest of the paper, we study how AMs using Q-learning algorithms set their prices in such an environment. We consider two cases. In the first case, analyzed in Section 3, we consider an

environment in which $\overline{\tau} = 1$ (a single trading round per episode). Our focus in this case is on whether and how outcomes when prices are set by AMs differ from those obtained in two benchmarks: (i) the Nash-Bertrand equilibrium with multiple dealers (the competitive case) and (ii) the case in which dealers set their price to maximize their aggregate expected profit (the monopoly case). This comparison will help us to analyze how AMs exert market power and cope with adverse selection relative to rational Bayesian dealers. In the second case, analyzed in Section 4, we consider a dynamic environment in which $\overline{\tau} = 2$ (two trading rounds per episode) and we focus on price discovery (i.e., on how AMs adjust their quotes over time).

3 The Static Case ($\bar{\tau} = 1$)

In this section, we compare the pricing policies chosen by AMs using a Q-learning algorithm to equilibrium outcomes predicted by standard economic analysis in the environment described in Section 2 when there is a single trading round per episode. We refer to this case as the static case since, when we solve for dealers' equilibrium pricing policies in this case, they behave as if they were facing a static one-shot problem. We proceed in three steps. First, in Section 3.1, we derive the equilibrium outcomes in two benchmark cases (the monopolist and competitive cases). Then, in Section 3.2, we describe the Q-learning algorithms used by AMs to choose their pricing policy when $\bar{\tau} = 1$. Third, in Section 3.3, we compare the pricing policies chosen by AMs to those obtained in the benchmarks.

3.1 Benchmarks

Monopolist Case. In the monopolist case, in each episode, each dealer chooses her price, denoted a^m, to maximize dealers' aggregate expected profit. Recalling that $Pr(\tilde{v} = v_H) = \mu = 1/2$, a^m solves:

$$a^m \in arg\max_a \bar{\Pi}\left(a, \frac{1}{2}\right). \tag{6}$$

Economic theory predicts that this price should be the equilibrium outcome when $N = 1$.

Competitive Case. In the competitive case, dealers choose a price a^c such that each dealer's

expected profit is nil. That is,

$$a^c \ s.t. \ \bar{\Pi}\left(a^c, \frac{1}{2}\right) = 0. \tag{7}$$

When the set of prices is continuous, a^c is the Bertrand-Nash equilibrium of the game played by dealers in each trading round. This is the outcome predicted by economic theory when $N \geq 2$.

We explain how to obtain a^c and a^m in Appendices A.4 and A.3 and we find (numerically) that (i) the competitive price, a^c, increases with Δv and decreases with σ while (ii) the monopoly price increases with both Δv and σ. We provide a numerical example in Table 1 where we report a^m and a^c when $\Delta v = 4$ and $\sigma = 5$ ($v_H = 4$ and $v_L = 0$, so that $\mathbb{E}(\tilde{v}) = 2$), the baseline values of the parameters in our experiments. The table also reports the expected half-quoted spread, $\mathbb{E}(a - \mathbb{E}(\tilde{v})) = a - \mathbb{E}(\tilde{v})$, (the difference between the ask price posted by each dealer and the unconditional expected payoff of the asset) and the expected half realized spread, $\mathbb{E}(a - \tilde{v} \mid V = 1)$. In contrast to the average half quoted spread, the expected half realized spread measures the expected profit of a dealer conditional on a trade by the client. As this trade is more likely when v is high, this measure accounts for the adverse selection cost borne by the dealer. In fact the difference between average half quoted spread and average half realized spread is often used as a measure of adverse selection costs in empirical studies.[11] Last observe that the total expected profit of a dealer is the expected half realized spread times the probability that the client trades ($\bar{\Pi}(a, \mu) = \Pr(V = 1)\mathbb{E}(a - \tilde{v} \mid V = 1)$).

[Insert Table 1 about here]

When the dispersion of clients' private valuations (σ) increases or the volatility of the asset payoff (Δv) decreases, dealers' ask prices become lower in the competitive case because dealers' adverse selection costs decline. In contrast, when the dispersion of clients' private valuations increases, the monopolist offer becomes larger, despite the fact that their adverse selection costs decline. This reflects an increase in dealers' rents, as shown by the increase in the expected realized spread. The reason is that as σ increases, clients' demand becomes more inelastic, which as usual enables a monopolistic dealer to extract larger rents. In contrast, when Δv increases, the client's demand becomes more elastic and the adverse selection cost increases. As a result, the monopolist dealer

[11] See Foucault *et al.* (2013), ch. 2, for a description of various measures of bid-ask spreads in securities markets and their interpretation.

charges a larger price but she obtains smaller rents (the realized bid-ask spread declines).

3.2 Q-Learning Algorithms

3.2.1 Description of the Algorithms

We now describe the functioning of Q-learning algorithms in the environment described in Section 2 when $\bar{\tau} = 1$.[12] We use the same notations as in Section 2, unless otherwise stated. In contrast to the benchmark case, we restrict AMs to choose their quotes in a discrete and finite action set $\mathcal{A} = \{a_1, a_2 ... a_M\}$, where each a_m is a possible ask price.[13]

To each dealer n and episode t, we associate a so-called *Q-Matrix* $\mathbf{Q}_{n,t} \in \mathbb{R}^{M \times 1}$. In this section, $\mathbf{Q}_{n,t}$ is simply a column vector of size M. The m-th entry of the matrix is denoted $q_{m,n,t}$ where $q_{m,n,t}$ represents the estimate in episode t of the payoff that AM n expects from playing price a_m. The Q-learning algorithm is meant to refine the payoff estimates in $\mathbf{Q}_{n,t}$ over time, and to end up playing the action associated with the highest estimate.

More formally, the algorithms (AMs) play the game according to the following process. We first initialize the matrices $\mathbf{Q}_{n,0}$ with random values: Each $q_{m,n,0}$ for $1 \leq m \leq M$ and $1 \leq n \leq N$ is i.i.d. and follows a uniform distribution over $[\underline{q}, \overline{q}]$. Then, in each episode t, we do the following:

1. For each dealer n, we define $m^*_{n,t} = arg\max_m q_{m,n,t-1}$ the index associated with the highest value in matrix $\mathbf{Q}_{n,t-1}$, and we denote $a^*_{n,t} = a_{m^*_{n,t}}$ the *greedy price* of AM n in episode t. This is the price that seems to maximize the AM's static profit, according to the estimates available in episode t.

2. For each dealer n, with probability $\epsilon_t = e^{-\beta t}$ the AM "explores" by playing $a_{n,t} = a_{\tilde{m}_{n,t}}$, where $\beta > 0$ and $\tilde{m}_{n,t}$ is a random integer between 1 and M, all values being equiprobable. The price $a_{\tilde{m}_{n,t}}$ is thus a price taken randomly in $\mathcal{A}$. With probability $1 - \epsilon_t$, the dealer "exploits" and plays $a_{n,t} = a^*_{n,t}$, the greedy price. The random draws leading to exploring or exploiting are i.i.d. across all dealers in a given episode.

[12] See Calvano *et al.* (2020) for an introduction to Q-learning algorithms in the more complex case of infinite horizon problems. See also Sutton and Barto (2018) for an introductory textbook on this topic.

[13] This constraint is necessary because the algorithm must evaluate the payoff associated with each possible price.

3. We compute $a_t^{min} = \min_n a_{n,t}$ the *best ask* in episode t, z_t the number of AMs with $a_{n,t} = a_t^{min}$, and draw $\tilde{v}_t$ and $\tilde{L}_t$. Each dealer n then receives a profit equal to $\pi_{n,t} = \Pi(a_{n,t}, a(t), \tilde{L}_t, \tilde{v}_t)$, as given by (2).[14]

4. We update the Q-matrix of each dealer as follows, with $0 < \alpha < 1$:

$$\forall 1 \leq n \leq N, q_{m,n,t} = \begin{cases} \alpha\pi_{n,t} + (1-\alpha)q_{m,n,t-1} & \text{if } a_{n,t} = a_m \\ q_{m,n,t} & \text{if } a_{n,t} \neq a_m \end{cases} \tag{8}$$

5. We then repeat starting from stage 1, until the last episode T.

Intuitively, each Q-learning algorithm alternates between experimenting random prices, and playing the price that seems to lead to the highest payoff based on past plays. As the number of past episodes grows, information accumulates and there should be less value in experimenting. For this reason, the probability of experimenting decays over time (here exponentially, at rate β). This important parameter of the algorithm governs the trade-off between experimenting and exploiting. A second trade-off is how much should one react to one particular observation $\pi_{n,t}$, knowing that payoffs are stochastic. This is governed by the parameter α: if α is large the algorithm reacts quickly to new observations, but the estimates generated in the Q-matrix are unstable (consider the extreme case $\alpha \to 1$). Conversely, if α is small the estimates are stable but it will take a lot of experimentation to move the values of the Q-matrix towards accurate estimates of the expected payoffs associated with each price.

3.2.2 Convergence

There are many variants of the Q-learning algorithm, with different specifications for the experimentation probability ϵ_t and the updating rule (8). The one described in the previous section is common in practical applications and is also the one used in recent papers in the economic literature (e.g., Calvano *et al.* (2020)). We choose it for comparability with prior literature. This version of Q-learning does not satisfy the assumptions given in, e.g., Watkins and Dayan (1992),

[14] We index all variables by the episode counter and omits the trading round index, τ within an episode since $\tau = 1$ in each episode here.

Jaakkola *et al.* (1994), or Tsitsiklis (1994) to guarantee convergence. In fact, given the design of these algorithms and the environment in which they operate, Lemma 1 below shows that no matter t (that is, even when T becomes very large), with a probability that is bounded away from 0, the Q-matrix changes by an amount that is bounded away from 0. Thus, entries in the Q-matrix never converge.

Lemma 1. (Impossibility of convergence of the Q-matrix) *For any given t and $a_m \in \mathcal{A}$, if $a_{n,t} = a_m = a_t^{\min}$, then ,*

$$\Pr\left(|q_{m,n,t} - q_{m,n,t+1}| \geq \Delta_m^*\right) \geq P_m^*,$$

where

$$\Delta_m^* := \frac{\alpha}{2}\left(v_H - v_L + \left|a_m - \frac{v_H - v_L}{2}\right|\right),$$

and

$$P_m^* := \min\left\{\frac{1}{2N}D(a_m, v_L), 1 - \frac{1}{2}(D(a_m, v_L) + D(a_m, v_H))\right\}$$

For instance, consider the case $N = 1$ and suppose that the AM plays for T consecutive periods the same price a_m. Then, as T goes to infinity, the value of the Q-matrix $q_{m,t}$ does not converge in probability to $\overline{\Pi}(a_m, \mu)$, that is the actual monopolist dealer's expected profit when she sets a price of a_m (the metrics a monopolist would use to set his price in economic theory). Because the Q-matrix does not converge, the price that maximizes the Q-matrix needs not stay the same and will vary with probability 1 after sufficiently many episodes. Thus, one cannot expect the greedy price to remain stable, even when T becomes very large. This also implies that the greedy price observed in the final episode will vary across experiments.

In actual simulations, when α is small, most experiments give the impression of "converging" in the sense that, after a sufficiently large number of episodes, the price chosen by each AM stays constant for many periods (this is because Δ_m^* is linear in α). This is the meaning of "convergence" in many papers in the literature (e.g., Calvano *et al.* (2020)). Thus, following the literature, we say that an experiment has "converged" if all algorithms' actions have been constant over the last κT periods (e.g., $\kappa = 0.05$). Moreover, to describe the outcome of the interaction between AMs

we look at the distribution of prices after a large number T of episodes and across a large number K of experiments, focusing in particular on the mean of the distribution. When needed, we use a superscript k to denote the outcome of the k^{th} experiment. For instance, $a_{n,t}^{*k}$ is the greedy price of dealer n in episode t of experiment k.

Keeping these observations in mind, we set the parameters of our baseline simulations as follows. The parameters of the economic environment are the same as in Table 1: $\Delta v = 4$, $\sigma = 5$, $v_H = 4$, and $v_L = 0$. In addition, the set of available prices $\mathcal{A}$ is all integers between 1 and 15 included. We initialize the Q-matrices with values between $\underline{q} = 3$ and $\overline{q} = 6$ so that all values of the initial Q-matrix are above the maximal payoff a dealer can get in a given period.[15] There are $K = 10,000$ experiments, $T = 200,000$ episodes per experiment, and in all experiments we set $\beta = 0.0008$ and $\alpha = 0.01$. This means that the algorithm chooses to experiment 1249.5 times in expectation, and hence "tries" each price about 100 times on average. As $\alpha = 0.01$, this frequency of experimentation is enough (in expectation) to override the initial values of the Q-matrix.[16] Finally, we set $\kappa = 0.05$, so that an experiment is said to "converge" if algorithms' actions have been unchanged for the last 10,000 episodes.

3.3 Results

In this section, we report the main results of our experiments. We first consider the monopoly case ($N = 1$) and duopoly case ($N = 2$), holding other parameters to their baseline values (Sections 3.3.1 and 3.3.2). In particular, we compare the distribution of final prices in these cases to their equilibrium values when $N = 1$ (monopoly price) and $N = 2$ (duopoly case), accounting for the fact that dealers must position their prices on a grid. Given this constraint, the theoretical monopoly price is $a^m = 7$ and there are two possible Nash-Bertrand equilibria ($a^c = 3$ or $a^c = 4$).

[15] This specification is common in the literature on Q-learning to guarantee that all actions are chosen sufficiently often to overcome the initial values of the Q-matrix. Indeed, as long as $q_{m,n,t}$ is larger than the maximal payoff the agent can realize, action m will necessarily be picked again because all the cells associated with actions that are played eventually fall below the maximal payoff.

[16] Note that Q-learning algorithms are meant for situations in which agents have no prior knowledge of the environment. Hence, there is no basis on which one could optimize the algorithm, e.g., by picking the "best" values of α and β. Rather, these values and the rules used by the algorithm must be seen as parameters.

3.3.1 A single AM ($N = 1$) behaves more competitively than in theory

Consider the case in which $N = 1$ first. Panel (a) of Figure 1 reports the evolution of the greedy price $a_{1,t}^{*k}$ over episodes, averaged over the $10,000$ experiments while Panel (b) reports the distribution over the K experiments of the final greedy price, $a_{1,T}^{*k}$ (whether convergence takes place or not). Panel (a) suggests that, on average the greedy price converges as the number of episodes becomes large. However, only 73.64% of the experiments converge (as defined in Section 3.2.2) and the final greedy price is heterogeneous across experiment as shown in Figure 1. The average final greedy price, $a_{1,T}^{*k}$, across all experiments is 6.16. However, there is substantial heterogeneity across experiments. As panel (b) shows, while most experiments ultimately reach a greedy price of 6, a substantial number reach 5 or 7, and in a few cases even 8 or 9. This dispersion in final outcomes across experiments is due to the environment being stochastic. Even though the values of the Q-matrix are not very sensitive to individual observations (remember that $\alpha = 0.01$), there is still a significant probability to obtain sufficiently many "bad draws" resulting in zero demand with prices of 6 or 7 to lead to a Q-matrix with a greedy price of 5 or 8, even though the monopolist's payoff is maximized at 7.

[Insert Fig. 1 here.]

A striking feature of this experiment is that, most of the times (in more than 75% of all experiments), the algorithm fails to learn the theoretical (optimal) monopoly price 7 even though T is large. Moreover, this failure is not random: On average, the final price posted by the algorithm is below 7. The modal price is 6, and the algorithm is more likely to set a price of 5 than a price of 8, even though playing 8 would give a higher expected profit than playing 5. The reason is that the updating rule (8) is biased against actions giving a high payoff with a low probability, such as choosing a high price. This effect is more pronounced when α is larger, but still significant even with the low value of α we are using (in unreported results, we checked that the average final greedy price indeed decreases in the parameter α).[17]

[17] To understand this point, imagine there are only two actions a_1 and a_2. Action a_1 gives a sure payoff π_1, whereas a_2 gives a payoff $\pi^+ > \pi_1$ with probability p, and $\pi^- < \pi_1$ with probability $1-p$, with $\pi_2 = p\pi^+ + (1-p)\pi^- > \pi_1$. If both actions are played many times, the expectation of $q_{2,t}$ associated with a_2 will converge to π_2. However, as noted in Lemma 1, the random variable $q_{2,t}$ itself does not converge pointwise. Instead, the distribution of $q_{2,t}$ converges to a non-degenerate distribution. A simple example is the case $\alpha = 1$: then $q_{2,t}$ will be equal to $\pi^+ > \pi_1$ with probability p and $\pi^- < \pi_1$ with probability $1-p$. Hence, the Q-learning algorithm will mistakenly pick action 1 as the greedy action with probability $1-p$.

One might be tempted to interpret this failure to learn the optimal price with probability 1 as a deficiency of the algorithm. However, this class of algorithms is not explicitly designed to learn the optimal price. Rather they seek to reach a certain balance between "exploring" and "exploiting". For instance, one could reach a final outcome closer to the monopoly price by choosing smaller values of α and β. However, doing so would be at the cost of playing suboptimal prices for more periods (so that the average profit of the AMs over all episodes might be smaller).

In any case, an important conclusion from the single dealer case is that the Q-learning algorithm used by the AMs in our experiments is not by itself biased towards high prices. If anything, the single dealer case shows that the opposite happens. This makes the non competitive final outcomes observed in the duopoly case (see next section) more striking.

3.3.2 Two AMs do not suffice to obtain Bertrand-Nash outcomes

Now consider the duopoly case ($N = 2$). The starting values of the Q-matrices, $\mathbf{Q}_{1,0}$ and $\mathbf{Q}_{2,0}$, for each AM as well as all the subsequent random draws for the two AMs, are drawn independently of each other (except the client's demand). Panel (a) of Figure 2 reports the evolution of the greedy price $a_{n,t}^{*k}$ for each AM over T episodes, averaged over the K experiments. Panel (b) reports the distribution of the final greedy price $a_{n,T}^{*k}$ for each AM over the K experiments.

[Insert Fig. 2 here.]

As can be seen in Figure 2, the AMs' quotes converge more quickly in the duopoly case than in the monopoly case. Convergence is also more frequent: In 94.18% of the experiments, the quote posted by each AM has converged after 200,000 episodes (vs., only 73.64% when $N = 1$). Moreover, in all experiments with convergence, the AMs end up posting the same price ($a_{1,T}^{*k} = a_{2,T}^{*k}$). However, this price is rarely one of the two Bertrand-Nash equilibrium prices (3 or 4 due to price discreteness). Indeed, we observe $a_{1,T}^{*k} = a_{2,T}^{*k} = 4$ in 5.57% of experiments only, and we never observe $a_{1,T}^{*k} = a_{2,T}^{*k} = 3$. As Panel (b) of Figure 2 shows, a majority of experiments (more than 60%) converge to a price of 5, about 20% converge to a price of 6, and some to 7 (the monopoly price) or 8. Thus, on average, the prices posted by the two competing AMs are far above the Bertrand-Nash equilibrium price.

The reason for this seemingly collusive outcome is different from the one in Calvano *et al.* (2020), because our setup precludes dynamic strategies (quotes cannot be contingent on past competitors' quotes in our set-up). Its origin seems closer to that in Asker *et al.* (2022) who also find that prices set by Bertrand competitors using Q-learning are above competitive prices (in an environment without adverse selection). In the first episodes, both AMs are experimenting with a high probability. AM 1 for instance is gradually learning how to best respond to AM 2. However, most of the time, AM 2 chooses a random price since the likelihood of experimentation is high in early episodes. The best response to AM 2 is actually for AM 1 to play $a = 6$. As AM 1 plays 6 more and more often (since the likelihood of experimentation declines over time), AM 2 should in principle learn that her best response is then to play $a = 5$ (in an undercutting process typical of Bertrand competition). However, because both AMs experiment less and less often over time, this undercutting process will typically not last long enough to reach the Bertrand outcome. For instance, both AMs may have reached a price of only 5 when the probability of experimenting ever again becomes very small. If for both AMs playing 3 or 4 did not prove profitable in the past (when the other AM was playing differently), then the AMs appear "stuck" with supra-competitive prices.[18] Our next step is to study how the probability that this happens depends on the parameters of the model, and in particular on the degree of adverse selection.

3.3.3 Adverse selection tends to make AMs' quotes more competitive

In this section we study how the outcomes of the simulations vary when we change the parameters of the economic environment, in particular the degree of adverse selection. For each set of parameters, in each experiment k and episode t we compute the following four variables (which correspond to empirically observable quantities):

1. **The trading volume** V_t^k, which is equal to 1 if a trade happens and 0 otherwise.

2. **The quoted spread** QS_t^k, which is the best ask minus the asset's ex ante expected value:

$$QS_t^k = a_t^{\min,k} - \mathbb{E}[\tilde{v}]. \tag{9}$$

[18] See Abada *et al.* (2022) for a comprehensive analysis and discussion of this issue. Wunder *et al.* (2010) show that even in a simple prisoner's dilemma Q-learning algorithms may not reach the Nash equilibrium.

3. **The realized spread** RS_t^k, which is:

$$RS_t^k = a_t^{\min,k} - v_t^k. \tag{10}$$

The realized spread is computed only when there is a trade. It measures the profit actually realized by the AM with the best quote, given the actual value v_t^k of the asset. Its average value over trades is a standard measure of dealers' expected profits per share in the literature (see Section 3.1).

We then compute the average across the K experiments of these three quantities in the last episode. That is, we compute:

$$\overline{V} = \frac{\sum_{k=1}^{K} V_T^k}{K} \tag{11}$$

$$\overline{QS} = \frac{\sum_{k=1}^{K} QS_T^k}{K} \tag{12}$$

$$\overline{RS} = \frac{\sum_{k=1}^{K} V_T^k RS_T^k}{\sum_{k=1}^{K} V_T^k}. \tag{13}$$

$$\tag{14}$$

[Insert Fig. 3 here.]

Panels a) and b) in Figure 3 show the effect of a change in σ (the variance of clients' private valuation) and Δv (the volatility of the asset payoff) on the average trading volume, the average quoted spread, and the average realized spread in the case with a single AM (dashed line), two AMs (plain line) and in the Bertrand-Nash equilibria (dotted lines).

As explained previously, an increase in σ reduces dealers' exposure to adverse selection and the elasticity of clients' demand to dealers' price. In the benchmark case (see Table 1), the first effect reduces adverse selection costs. For this reason, the quoted spread in the Bertrand-Nash equilibria decreases (weakly due to price discreteness) with σ. However, surprisingly, the opposite pattern is observed for the quoted spread posted by AMs: As σ increases, the two AMs post less competitive quotes. In fact, the effect of σ on AM's quotes is similar to its effect on the monopoly price (red dashed line in 3). In this case, like a monopolist, the competing AMs seem to take advantage of

the decrease in the client's demand elasticity to charge larger markups and thereby obtain larger expected profits (as shown by the evolution of the average realized spread).[19]

Thus, surprisingly, a decrease in adverse selection makes the quotes posted by AMs less competitive. The effect of Δv on AMs' quotes (Panel b) conveys a similar message. As Δv decreases from 8 to 4, AMs' exposure to adverse selection decreases. However, as shown by the evolution of AMs' realized spread, their rents increase, exactly as in the monopolist case. When Δv keeps decreasing (from 4 to 1), AMs rents decrease but in a way similar to what is observed in theory for the monopolist.[20] In sum, competing AMs react to a decline in adverse selection (an increase in σ or a decrease in Δv) in a way qualitatively similar to a monopolist rather than Bertrand competitors.

Panel (c) of 3 shows the effect of an increase in the number of AMs (from 1 to 10). As the number of AMs increases, AMs' quotes become closer to the Bertrand-Nash equilibria. Thus, AMs' rents (realized bid-ask spreads) decline. This pattern may seem intuitive. However, in theory it takes only two dealers to obtain the Bertrand-Nash equilibrium. Thus, economic theory predicts that bid-ask spreads and dealers' rents should decline when N increases from 1 to 2 but that a further increase in N should have no effect. Empirical findings regarding the effects of high frequency market makers' entry on bid-ask spreads, reported in Brogaard and Garriott (2019) (discussed in the introduction), are more consistent with the patterns obtained for AMs than those predicted by the Bertrand-Nash equilibrium.

3.3.4 Welfare implications of algorithmic market-making

Spread measures do not immediately translate into welfare measures. In particular, the realized spread RS measures a market-maker's realized profit (and hence, cost for the client) conditionally on a trade, but does not take into account the probability that this trade occurs. To further investigate the consequences of AMs for total welfare in the economy and its distribution between market-makers and buyers, we compute the levels of welfare, consumer surplus, and firm profits achieved with AMs and compare them to their counterparts in the competitive benchmark.

[19] The decline in the client's demand elasticity explains why trading volume increases with σ in the experiments, despite the fact that AMs charge a larger price to their client.

[20] AMs' rents also evolve in a way similar to that observed in one of the two Nash Bertrand equilibria (dotted purple line) but opposite to that in the other one (yellow dashed line). We think that the pattern observed in first case is due to price discreteness and will therefore not be robust with a finer grid, in contrast to other patterns.

For a given best ask a, total welfare can be computed as:

$$W(a) = \Pr(\tilde{v} + \tilde{L} \geq a)\mathbb{E}[\tilde{L}|\tilde{v} + \tilde{L} \geq a]. \tag{15}$$

In words, welfare in this model is driven by the liquidity shocks $\tilde{L}$, which create gains from trade between buyers and market-makers. Welfare is always lower when the ask price increases, and as a result even in the competitive case as increase in adverse selection lowers welfare. Welfare can be further decomposed into consumer surplus CS and producer surplus PS:

$$\begin{aligned} CS(a) &= \Pr(\tilde{v} + \tilde{L} \geq a)\mathbb{E}[\tilde{L} + \tilde{v} - a|\tilde{v} + \tilde{L} \geq a], & (16) \\ PS(a) &= \Pr(\tilde{v} + \tilde{L} \geq a)\mathbb{E}[a - \tilde{v}|\tilde{v} + \tilde{L} \geq a]. & (17) \end{aligned}$$

Based on the results of the experiments, we compute the average realized values of W, CS, and PS, and show in Fig. 4 how they vary with Δv and σ.

[Insert Fig. 4 here.]

We observe that an increase in σ leads to an increase in profits, due to both the AMs behaving less competitively (realized spreads increase) and demand elasticity being lower. However, because this elasticity is low, high prices have a lower impact on the probability that a trade is realized, and conditionally on a trade the gains are also higher. As a result, consumer surplus and total welfare also increase with σ. An increase in Δv has a somewhat ambiguous impact on realized spreads but it reduces profits, consumer surplus, and hence total welfare.

Overall, the comparative statics of welfare and profit with respect to σ and Δv are the same in the two benchmarks and with a duopoly of AMs, the levels reached with AMs being in between the monopoly benchmark and the competitive benchmark.

4 Price Discovery ($\bar{\tau} = 2$)

Models of trading with asymmetric information in financial markets are often used to study the process by which market participants discover asset fundamental values ("price discovery"). In these models, trades convey information about an asset payoff (because some trades come from informed investors). Using this information, uninformed traders (e.g., dealers) update their beliefs about this payoff in a Bayesian way. Via this dynamic learning process, over time, prices get closer to the asset value (see, for instance, Glosten and Milgrom (1985) or Easley and O'Hara (1992)).

In this section, we study whether AMs can also discover asset fundamental values ($\tilde{v}$ in our setting). To do so, we consider the case with two trading rounds ($\bar{\tau} = 2$), following the same steps as when $\bar{\tau} = 1$. That is, in Section 4.1, we first explain how to derive equilibrium prices in our two benchmarks (the monopoly case and the Bertrand-Nash equilibrium). Then, we explain how Q-learning algorithms work in this environment (Section 4.2). Finally we present the results in Section 4.3.

4.1 Benchmarks: Learning the Fundamental Value

When $\bar{\tau} = 2$, dealers can learn information about $\tilde{v}$ from the trading outcome at date 1. Thus, their beliefs regarding the payoff of the asset evolve over time. As is standard in models of trading with asymmetric information, in the benchmark monopoly and competitive cases, we assume that dealers update their beliefs in a Bayesian way. At the end of the first trading round in a given episode, there are two possible trading histories (H_1): (i) a trade at price a_1^{min} ($H_1 = \{1, a_1^{min}\}$) or (ii) no trade ($H_1 = \{0, a_1^{min}\}$). In the first case, dealers' Bayesian beliefs about the likelihood that $v = v_H$ is (remember that dealers' prior belief about this event is 1/2):

$$\mu_2(1, a_1^{min}) := Pr(v = v_H \mid H_1 = \{1, a_1^{min}\}) = \frac{D(a_1^{min}, v_H)}{D(a_1^{min}, v_H) + D(a_1^{min}, v_L)}, \tag{18}$$

where $D(a, v)$, given by (4), is the probability that the client buys the asset at price a when the asset value if v. In the second case, dealers' Bayesian beliefs about the likelihood that $v = v_H$ is:

$$\mu_2(0, a_1^{min}) := Pr(v = v_H \mid H_1 = \{0, a_1^{min}\}) = \frac{1 - D(a_1^{min}, v_H)}{2 - (D(a_1^{min}, v_H) + D(a_1^{min}, v_L))}. \tag{19}$$

It is easily checked that $\mu_2(1, a_1^{min}) > \mu_2(0, a_1^{min})$ if (and only if) $\Delta v > 0$. That is, Bayesian dealers should revise their beliefs about the expected payoff of the asset upward after a trade (buy) at date 1 and downward after no trade at date 1.

Given these observations, one expects the monopoly price and the competitive price (the Nash-Bertrand equilibrium price) to be larger (smaller) in the second trading round than in the first if there is a trade (no trade) in the first trading round. Table 2 shows that this is the case for the parameters of our experiments. In addition, in the competitive case, the difference between dealers' ask prices when there is a trade and when there is no trade increases with the informativeness of the order flow in the first period (i.e., increases with Δv and decreases with σ). In addition, Table 2 shows that, in the competitive experiment, the ask price posted by dealers in the second period is smaller on average than in the first period (that is, $\mathbb{E}[a_2^c] - a_1^c \leq 0$). This reflects the fact that as time passes, the informational asymmetry between dealers and their clients decline since dealers learn information about the asset payoff. Thus, they face less adverse selection and therefore across all possible realizations of v and the trading history at date 1, their ask price should be closer to the asset unconditional value in the second period than in the first period. In Section 4.3, we study whether AMs' quotes satisfy these properties or not. This is a way to study whether AMs learn to discover the asset payoff, even though they are not programmed to be Bayesian, as competitive dealers do in the benchmark case.

[Insert Table 2 here.]

Table 2 also shows that, as in the case with one trading round (and for the same reasons), (i) the competitive and the monopoly prices increase with the volatility of the asset (Δv) in each trading round, (ii) the competitive prices in each trading round decrease with the dispersion of clients' private valuations (σ) and (iii) the monopoly prices in each trading round increase with this

dispersion.

Last, observe that, in the competitive case, the quotes posted by dealers in the first trading round are identical to those obtained when there is a single trading round (compare Tables 1 and 2). In contrast, the monopolist price in the first trading round differs from that obtained when there is a single a trading round. This reflects the fact that, in choosing her price in the first trading round, a monopolist accounts for the effect of this price on her expected trading profit in the first trading round *and* her expected trading profit in the second trading round via the effect of her choice on her belief about the asset payoff given the first period outcome (trade/no trade).[21]

4.2 Q-Learning Algorithms

In this section, we explain how we adapt the Q-learning algorithms described in Section 3.2 to the case in which episodes have two trading rounds. The algorithms will keep track in each episode of the "state" they are in, and will play an action depending on the state. More specifically, for each AM n, we define $(N+3)$ states, denoted s_n, as follows: (i) $s_n = \emptyset$ in the first trading round; (ii) $s_n = NT$ in the second trading round if no trade takes place in the first; (iii) $s_n \in \mathcal{S} = \left\{0, \frac{1}{N}, \frac{1}{N-1}, ..., \frac{1}{2}, 1\right\}$ is the number of shares sold by AM n if a trade took place in period 1 (depending on how many AMs shared the market). Each AM then relies on a Q-matrix $\mathbf{Q}_{n,t} \in \mathbb{R}^{M\times(N+3)}$, in which each line corresponds to a different price and each column to a state, ordered as in the previous paragraph. We denote $q_{m,s,n,t}$ the (m,s) entry of matrix $\mathbf{Q}_{n,t}$.

We then modify the process described in Section 3.2.1 as follows. For any experiment k, we initialize the matrices $\mathbf{Q}_{n,0}$ with random values: Each $q_{m,s,n,0}$ (for $1 \leq m \leq M$, $1 \leq n \leq N$, and $s \in \mathcal{S}$) is i.i.d. and follows a uniform distribution over $[\underline{q}, \overline{q}]$. Then, in each episode t, we do the following:

Period 1:

1. For each AM n, we define $m_{n,t}^{1,*} = arg\max_m q_{m,\emptyset,n,t-1}$ the index associated with the highest value in matrix $\mathbf{Q}_{n,t-1}$ in state $s = \emptyset$ (the first period), and we denote $a_{n,t}^{1,*} = a_{m_{n,t}^{1,*}}$ the

[21] One can show that $\mu_2(1, a_1^{min})$ and $\mu_2(0, a_1^{min})$ increase with a_1^{min}. Thus, by choosing a high a_1^{min}, the monopolist improves the informational content of a trade at date 1 but it reduces the informational content of observing no trade.

corresponding greedy price.

2. For each AM n, with probability $\epsilon_t = e^{-\beta t}$ the AM "explores" by playing $a^1_{n,t} = a_{\tilde{m}^1_{n,t}}$, where $\beta > 0$ and $\tilde{m}^1_{n,t}$ is a random integer between 1 and M, all values being equiprobable. With probability $1 - \epsilon_t$, the AM "exploits" and plays the greedy price $a^1_{n,t} = a^{1,*}_{n,t}$. The random draws leading to exploring or exploiting are i.i.d. across all AMs in a given trading round of a given episode.

3. We compute $a^{1,min}_t = \min_n a^1_{n,t}$, and draw $\tilde{v}_t$ and $\tilde{L}_{1,t}$. This determines the position $I^1_{n,t}$ taken by each AM in period 1 and the state $s_{n,t}$ it will be in when period 2 starts. Formally, denote $\mathcal{D}^1_t$ the set of AMs who quote $a^{1,min}_t$ and z^1_t the size of this set. Then, if $\tilde{v}_t + \tilde{L}_{1,t} \geq a^{1,min}_t$ we have $I^1_{n,t} = s_{n,t} = \frac{1}{z^1_t}$ for every $n \in \mathcal{D}^1_t$, and $I^1_{n,t} = s_{n,t} = 0$ for $n \notin \mathcal{D}^1_t$. If $\tilde{v}_t + \tilde{L}_{1,t} < a^{1,min}_t$ then $I^1_{n,t} = 0$ and $s_{n,t} = NT$ for every n.

4. We update the first column of the Q-matrix of each AM as follows:

$$\forall 1 \leq n \leq N, q_{m,\emptyset,n,t} = \begin{cases} \alpha[a^1_{n,t} I^1_{n,t} + \max_{m'} q_{m',s_{n,t},n,t-1}] + (1-\alpha) q_{m,\emptyset,n,t-1} & \text{if } a^1_{n,t} = a_m \\ q_{m,\emptyset,n,t-1} & \text{if } a_{n,t} \neq a_m \end{cases} \tag{20}$$

Period 2:

1. At the beginning of period 2 we know the state $s_{n,t}$ in which AM n finds itself. We define $m^{2,*}_{n,t} = arg\max_m q_{m,s_{n,t},n,t-1}$ the index associated with the highest value in matrix $\mathbf{Q}_{n,t-1}$ in state $s = s_{n,t}$, and we denote $a^{2,*}_{n,t} = a_{m^{2,*}_{n,t}}$ the corresponding greedy price.

2. With probability ϵ_t the AM plays a random price $a^2_{n,t}$, following the same process as in period 1. With probability $1 - \epsilon_t$, the AM plays $a^2_{n,t} = a^{2,*}_{n,t}$.

3. We compute $a^{2,min}_t = \min_n a^2_{n,t}$ and draw $\tilde{L}_{2,t}$. This determines the position $I^2_{n,t}$ taken by each AM in period 2, following the same rules as in period 1.

4. For each AM n, we only update the column corresponding to state $s_{n,t}$, as follows:

$$\forall 1 \leq n \leq N, q_{m,s_{n,t},n,t} = \begin{cases} \alpha[a^2_{n,t} I^2_{n,t} - \tilde{v}_t(I^1_{n,t} + I^2_{n,t})] + (1-\alpha) q_{m,s_{n,t},n,t-1} & \text{if } a^2_{n,t} = a_m \\ q_{m,s_{n,t},n,t-1} & \text{if } a^2_{n,t} \neq a_m \end{cases} \tag{21}$$

Q-learning algorithms were initially designed to solve dynamic stochastic optimization problems (both finite and infinite horizon), and are thus in principle well suited to optimizing prices in this environment. The Q-matrix is defined in such a way that each algorithm can in principle learn to play a different price in period 2 depending on the "state", that is, depending on whether there was a trade in period 1. Note that, in addition, the state needs to include the amount sold by the AM in period 1. Indeed, as v_t is only revealed in period 2, the Q-matrix can record only at the end of period 2 what was the actual cost of selling some units of the asset in period 1.[22]

4.3 Results

To study price discovery with AMs using the Q-learning algorithms described in the previous section, we proceed exactly as in Section 3.3. In particular, we use the same parameter values for K (number of experiments), T (number of episodes per experiments), α and β. For brevity, we only focus on the case with two AMs ($N = 2$). We measure price discovery by AMs (i.e., whether AMs' quotes reflect information about the asset payoff contained in the first period trade) by computing the magnitude of the average price reaction to the observation of a trade vs. no trade (across experiments with the same environment). Formally, defining $V_t^{\tau,k}$ the total trading volume in trading round τ of episode t in experiment k, we compute:

$$Discovery = \frac{\sum_{k=1}^K V_T^{1,k}[a_T^{2,min,k} - a_T^{1,min,k}]}{\sum_{k=1}^K V_T^{1,k}} - \frac{\sum_{k=1}^K (1 - V_T^{1,k})[a_T^{2,min,k} - a_T^{1,min,k}]}{\sum_{k=1}^K (1 - V_T^{1,k})}. \tag{22}$$

[22] Using inventory levels as the state variable is common in other applications of Q-learning, in particular in dynamic pricing and revenue management. See, e.g., Rana and Oliveira (2014) for a recent example. The list of states used by the algorithms is an important parameter of the model. The list could be even richer (e.g., conditioning on prices in period 1 as well), or coarser (not distinguishing states NT and 0).

The variable *Discovery* is the empirical counterpart, in our experiments, of the difference between the ask price in the second period when there is a trade and when there is no trade in the benchmark cases. In these cases, this difference is always positive because dealers become more optimistic about the asset payoff after observing a buy in the first trading round than after observing no buy (see Table 2).

We also want to study whether price discovery induces dealers to charge lower markups relative to their expectation of the asset payoff because it reduces informational asymmetries, as is observed when dealers are competitive in the benchmark case ($\mathbb{E}(a_2^c) < a_1^c$; see Table 2). To this end, we compute the average difference (denoted $Difference$) between the ask price posted in the second trading round and the ask price posted in the first trading round across experiments:

$$Difference = \frac{\sum_{k=1}^{K}[a_T^{2,min,k} - a_T^{1,min,k}]}{K}. \tag{23}$$

If AMs behave as in the competitive benchmark, $Difference$ should be negative. If it is not and $Discovery > 0$, this indicates that (i) price discovery takes place but (ii) AMs take advantage of the reduction in informational asymmetries to charge less competitive prices, in line with our observations in the static case.

Figure 5 plots $Discovery$ and $Difference$ for different values of σ and Δv. In addition, we plot the highest and lowest values these quantities can take across the several Nash equilibria of the game, the monopoly benchmark, and the competitive benchmark with continuous prices.

[Insert Fig. 5 here.]

First, we observe that for all values of the parameters, $Discovery$ is positive. Thus, AMs learn to quote higher prices when a trade occurred in period 1 than when a trade did not occur. Hence, Q-learning algorithms are able to learn from past trades and contribute to price discovery. However, the algorithms seem to significantly "overshoot". That is, the difference in the prices posted by AMs following a buy or no buy in the first trading round is always larger than that predicted in the most competitive Nash-Bertrand equilibrium (the dashed dotted line), given price discreteness. This indicates that the difference in posted prices following a trade or no trade in the first trading

round is in part driven by deviations from competitive prices.

Second, we observe that $Difference$ is always positive, that is on average the algorithms use a higher price in the second trading round than in the first. This is in stark contrast to the competitive benchmark in which, at least if the tick size were zero, $Difference$ should be negative (as shown in Table 2 and the dashed green line in Figure 5 in the case of "Difference").

A mechanism that explains both results is, as in the static case, that adverse selection curbs the market power of algorithms. In our set-up, observing the trading outcome in the first trading round always reduces informational asymmetries between dealers and clients in the benchmark case. Thus, dealers' adverse selection cost is smaller in the second trading round. This decrease in adverse selection leads the AMs to settle on using less competitive prices, as we already observed in the static case. In addition, adverse selection is reduced more after a trade than after no trade (observing a trade is less likely ex ante, hence is more informative when it happens). Thus, AMs tend to charge larger markups after a trade than after no trade, explaining why on average $Difference$ is positive insead of negative as in the competitive case.

These results give interesting insights into how competition between algorithms can be spotted in the data. The first result implies that quotes will tend to over-react to order flow, potentially generating more long-term reversal. The second result implies that spreads tend to widen as adverse selection is resolved over time, whereas in competitive environments the opposite should occur (see, e.g., Glosten and Putnins (2020)).

5 Conclusion

We study the interaction of market-makers using Q-learning algorithms in a standard microstructure environment a la Glosten and Milgrom (1985). We show that this provides a natural workhorse model to study the role of algorithms in securities markets, and how their behavior may differ from what is predicted by standard theory. We find that, despite their simplicity and the challenge of an environment with adverse selection, algorithms behave in a realistic way: their quotes reflect adverse selection costs and they update their quotes in response to the observed order flow. However, their behavior is markedly different from what standard theory predicts. In particular, their quotes tend

to be above the competitive level, and to become less competitive over time as adverse selection gets resolved. More generally, our analysis shows that the interaction between algorithms is significantly affected by the presence and extent of adverse selection, suggesting that securities markets are a quite specific and particularly interesting application of recent research on competition between algorithms.

References

Abada, I., Lambin, X. and Tchakarov, N. (2022). *Collusion by Mistake: Does Algorithmic Sophistication Drive Supra-Competitive Profits?* Working paper. 6, 17

Asker, J., Fershtman, C. and Pakes, A. (2021). *Artificial intelligence and pricing: The impact of algorithm design*. Tech. rep., National Bureau of Economic Research. 6

—, — and — (2022). Artificial intelligence, algorithm design, and pricing. *AEA Papers and Proceedings*, **112**, 452–56. 17

Baldauf, M. and Mollner, J. (2020). High-frequency trading and market performance. *The Journal of Finance*, **75** (3), 1495–1526. 6

Banchio, M. and Skrzypacz, A. (2022). Artificial intelligence and auction design. *Available at SSRN 4033000 9*. 6

Biais, B., Foucault, T. and Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, **116** (2), 292–313. 6

Brogaard, J. and Garriott, C. (2019). High-frequency trading competition. *Journal of Financial and Quantitative Analysis*, **54** (4), 1469–1497. 4, 19

—, Hagströmer, B., Nordén, L. and Riordan, R. (2015). Trading fast and slow: Colocation and liquidity. *The Review of Financial Studies*, **28** (12), 3407–3443. 1

Budish, E., Cramton, P. and Shim, J. (2015). The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response *. *The Quarterly Journal of Economics*, **130** (4), 1547–1621. 6

Calvano, E., Calzolari, G., Denicolo, V. and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, **110** (10), 3267–97. 5, 11, 12, 13, 17

Cartea, Á., Chang, P., Mroczka, M. and Oomen, R. C. (2022a). *AI driven liquidity provision in OTC financial markets*. Working paper. 6

—, — and Penalva, J. (2022b). *Algorithmic Collusion in Electronic Markets: The Impact of Tick Size*. Working paper. 6

Chen, L., Mislove, A. and Wilson, C. (2016). An empirical analysis of algorithmic pricing on amazon marketplace. In *Proceedings of the 25th international conference on World Wide Web*, pp. 1339–1349. 1

CMA (2018). Pricing algorithms. pp. 3–62. 1

Easley, D. and O'Hara, M. (1992). Time and the process of security price adjustment. *The Journal of Finance*, **47** (2), 577–605. 21

Foucault, T., Marco, P. and Ailsa, R. (2013). *Market Liquidity: Theory, Evidence, and Policy*. Oxford: Oxford University Press. 10

Glosten, L. and Putnins, T. (2020). *Welfare Costs of Informed Trade*. Working paper. 27

Glosten, L. R. and Harris, L. E. (1988). Estimating the components of the bid/ask spread. *Journal of financial Economics*, **21** (1), 123–142. 5

— and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, **14** (1), 71–100. 1, 3, 4, 6, 21, 27

Hansen, K. T., Misra, K. and Pai, M. M. (2021). Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science*, **40** (1), 1–12. 6

Hendershott, T., Jones, C. M. and Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, **66** (1), 1–33. 4

Jaakkola, T., Jordan, M. I. and Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, **6** (6), 1185–1201. 13

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, **53** (6), 1315–1335. 1, 3, 6

MacKay, A. and Weinstein, S. (2022). *Dynamic Pricing Algorithms, Consumer Harm, and Regulatory Response*. Working paper. 1

Menkveld, A. and Zoican, M. (2017). Need for speed? exchange latency and liquidity. *Review of Financial Studies*, **30** (4), 1188–1228. 6

Menkveld, A. J. (2013). High frequency trading and the new market makers. *Journal of financial Markets*, **16** (4), 712–740. 1

OECD (2017). Algorithms and collusion: Competition policy in the digital age. pp. 1–72. 1

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, **116** (2), 257–270. 6

Rana, R. and Oliveira, F. S. (2014). Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning. *Omega*, **47**, 116–126. 25

Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction*. Cambridge (Mass.): MIT Press. 11

Tsitsiklis, J. (1994). Asynchronous stochastic approximation and q-learning. *Machine Learning*, **16**, 185–202. 13

Watkins, C. and Dayan, P. (1992). Q-learning. *Machine Learning*, **8**, 279–292. 12

Wunder, M., Littman, M. L. and Babes, M. (2010). Classes of multiagent q-learning dynamics with epsilon-greedy exploration. In *ICML*, pp. 1167–1174. 6, 17

A Appendix

A.1 Tables

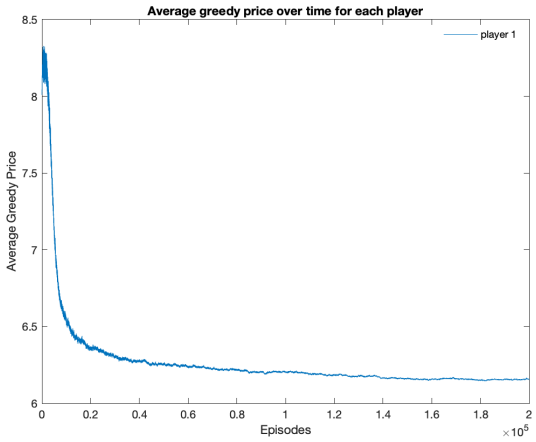
Panel A					
σ	0.5	1	3	5	7
Competitive Case					
a^c	4.00	4.00	3.24	2.68	2.47
Quo. Spread	2.00	2.00	1.24	0.68	0.47
Real. Spread	0	0	0	0	0
Monopoly					
a^m	4.37	4.69	5.68	6.54	7.03
Quo. Spread	2.37	2.69	3.68	4.54	5.03
Real. Spread	0.03	0.09	0.32	0.68	1.03
Panel B					
Δv	0	2	4	6	8
Competitive Case					
a^c	2	2.16	2.68	3.65	5.02
Quo. Spread	0	0.16	0.68	1.65	3.02
Real. Spread	0	0	0	0	0
Monopoly					
a^m	5.75	5.94	6.54	7.66	9.11
Quo. Spread	3.75	3.94	4.54	5.66	7.11
Real. Spread	0.82	0.78	0.69	0.57	0.47

Table 1: Predicted Outcomes in the Benchmark Cases, $\bar{\tau} = 1$. Prices are continuous. Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$ and $v_L = 0$). Panel A: $\Delta v = 4$. Quotes have been rounded up to two decimals (which explains why they are equal when $\sigma = 0.5$ and $\sigma = 1$). Panel B: $\sigma = 5$.

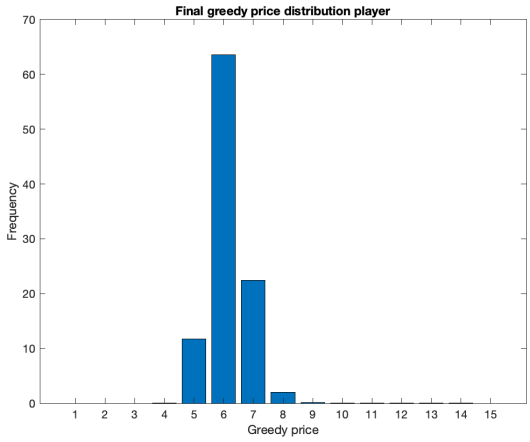
Panel A					
σ	0.5	1	3	5	7
Competitive Case					
a_1^c	4.00	4.00	3.24	2.68	2.47
$a_2^c, V_1 = 1$	4.00	4.00	3.82	3.26	2.92
$a_2^c, V_1 = 0$	4.00	4.00	2.44	2.08	2.02
$\mathbb{E}(a_2^c)$	4.00	4.00	2.96	2.62	2.45
Monopoly					
a_1^m	4.38	4.75	5.65	6.53	7.8
$a_2^m, V_1 = 1$	4.38	4.75	6.2	7.33	8.47
$a_2^m, V_1 = 0$	4.38	4.75	5.45	6.28	7.59
$\mathbb{E}(a_2^m)$	4.38	4.75	5.65	6.53	7.8
Panel B					
Δv	0	2	4	6	8
Competitive Case					
a_1^c	2	2.16	2.68	3.65	5.03
$a_2^c, V_1 = 1$	2	2.5	3.26	4.6	5.87
$a_2^c, V_1 = 0$	2	1.8	2.08	2.45	3.67
$\mathbb{E}(a_2^c)$	2	2.09	2.62	3.42	4.61
Monopoly					
a_1^m	5.76	5.94	6.53	7.61	9.09
$a_2^m, V_1 = 1$	5.76	6.2	7.33	8.61	9.73
$a_2^m, V_1 = 0$	5.76	5.87	6.28	7.26	8.86
$\mathbb{E}(a_2^m)$	5.76	5.93	6.49	7.53	9.01

Table 2: Predicted Outcomes in the Benchmark Cases, $\bar{\tau} = 2$. Prices are continuous. Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$ and $v_L = 0$). Panel A: $\Delta v = 4$. Quotes have been rounded up to two decimals (which explains why they are equal when $\sigma = 0.5$ and $\sigma = 1$). Panel B: $\sigma = 5$. In each case, $I_1 = 1$ if a trade takes place at date 1 and $I_1 = 0$ otherwise.

A.2 Figures

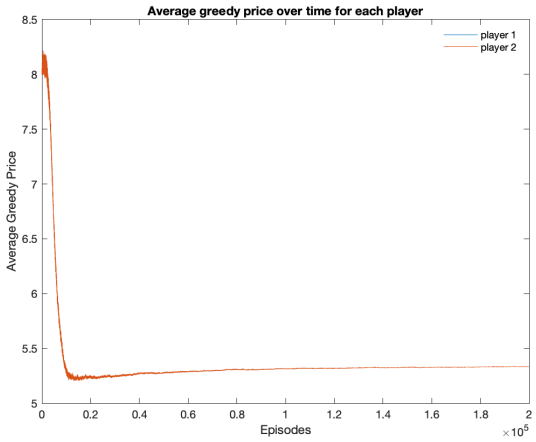


(a) Average greedy price as a function of time.

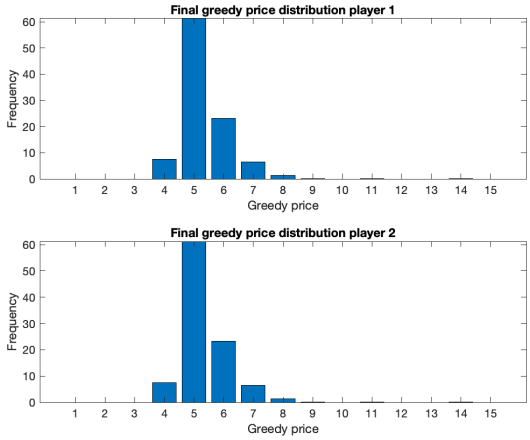


(b) Distribution of the final greedy price.

Figure 1: A single AM - Baseline Parameters.
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

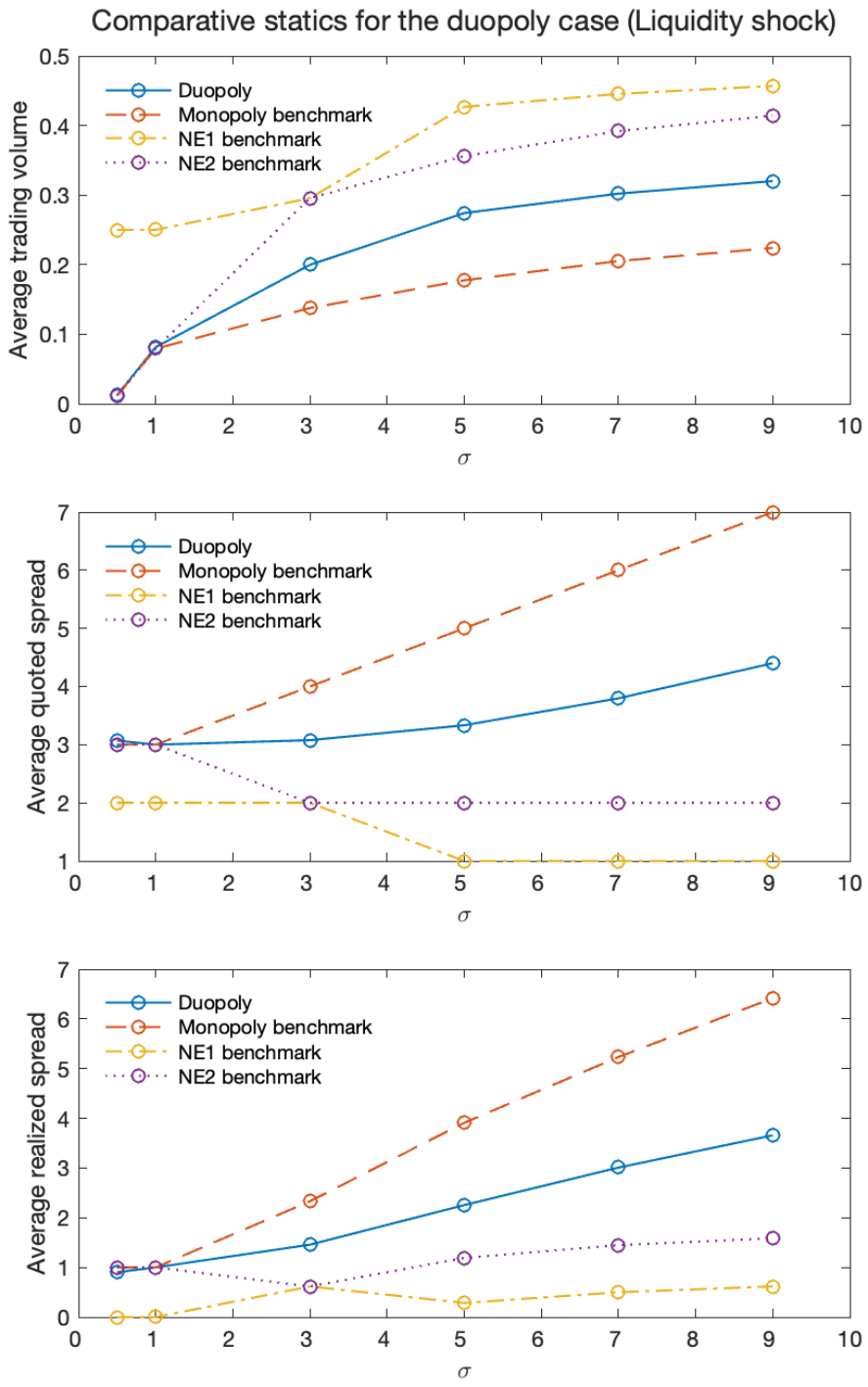


(a) Average greedy price of both AMs as a function of time.

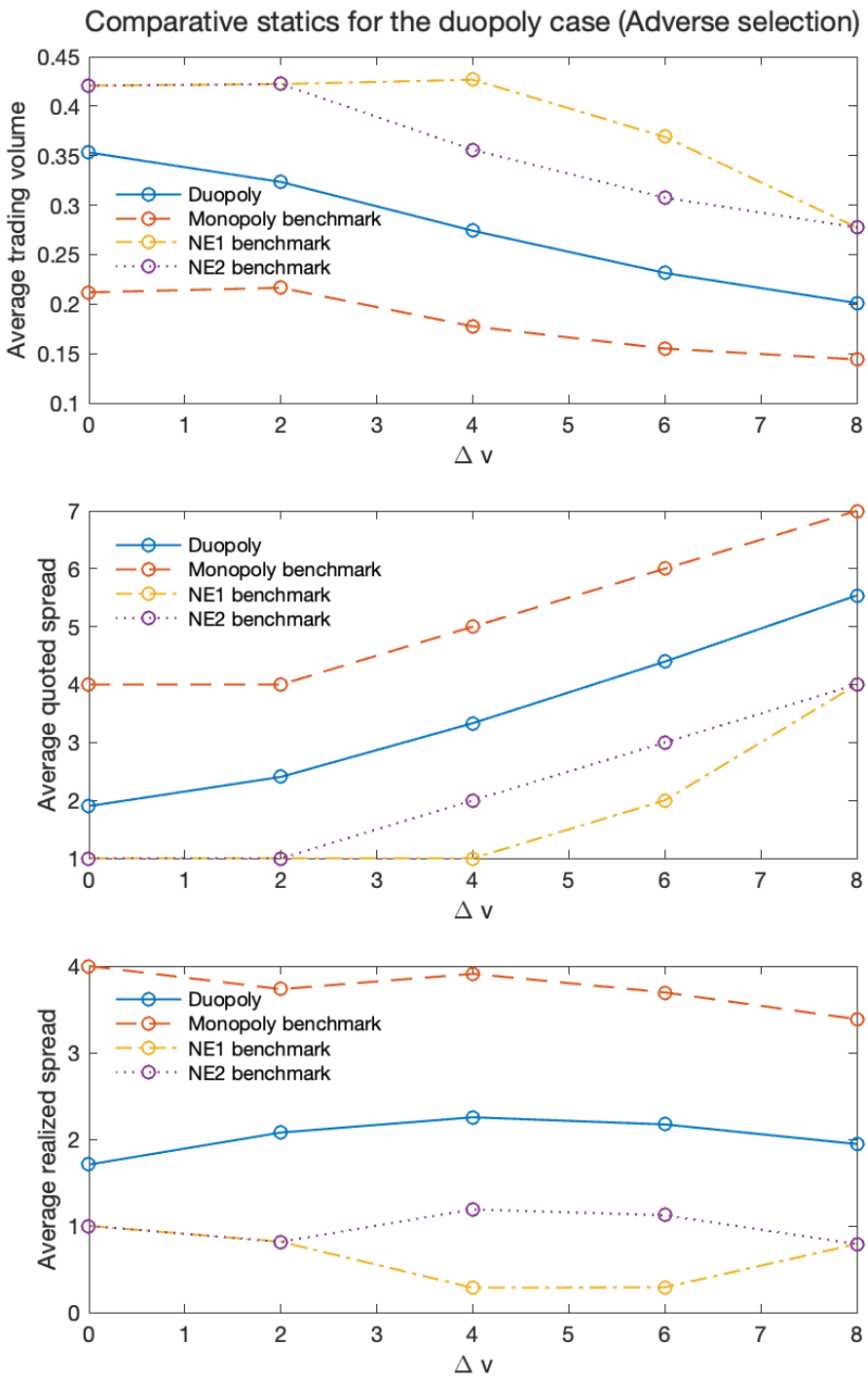


(b) Distribution of the final greedy price of both AMs.

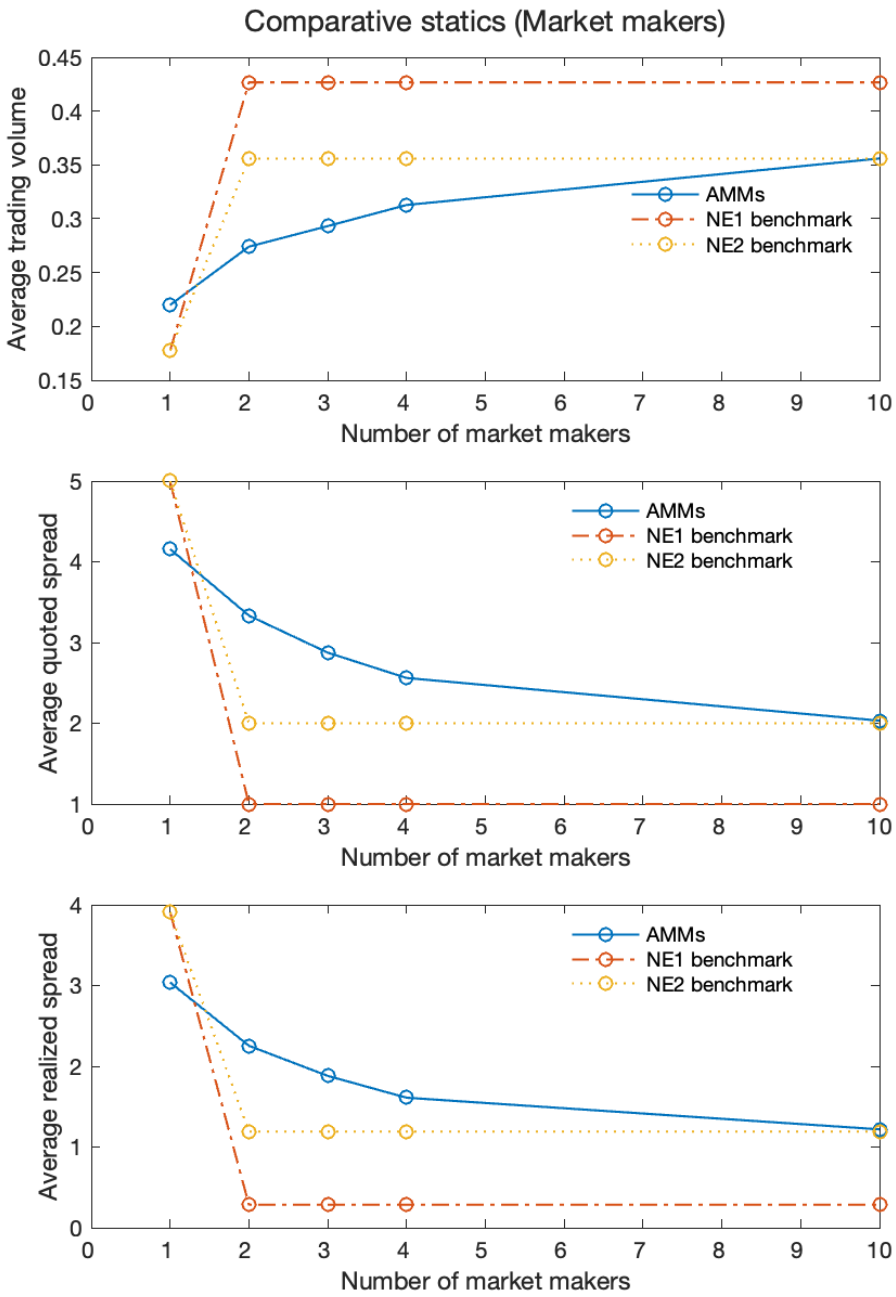
Figure 2: Duopoly of AMs - Baseline Parameters.
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)



(a) Dispersion of Clients' Private Valuations (σ)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

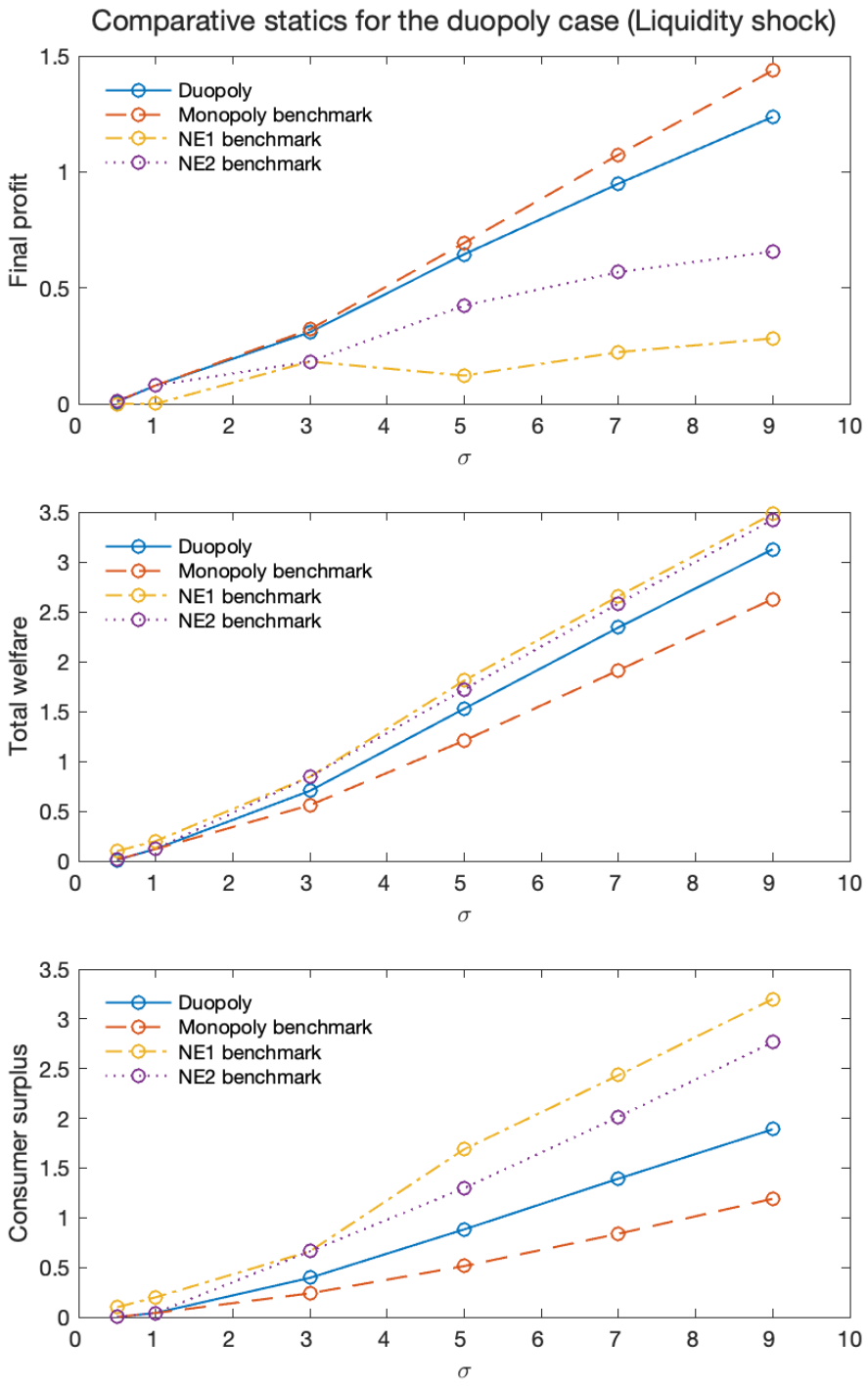


(b) Volatility of the Asset Payoff (Δv)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$

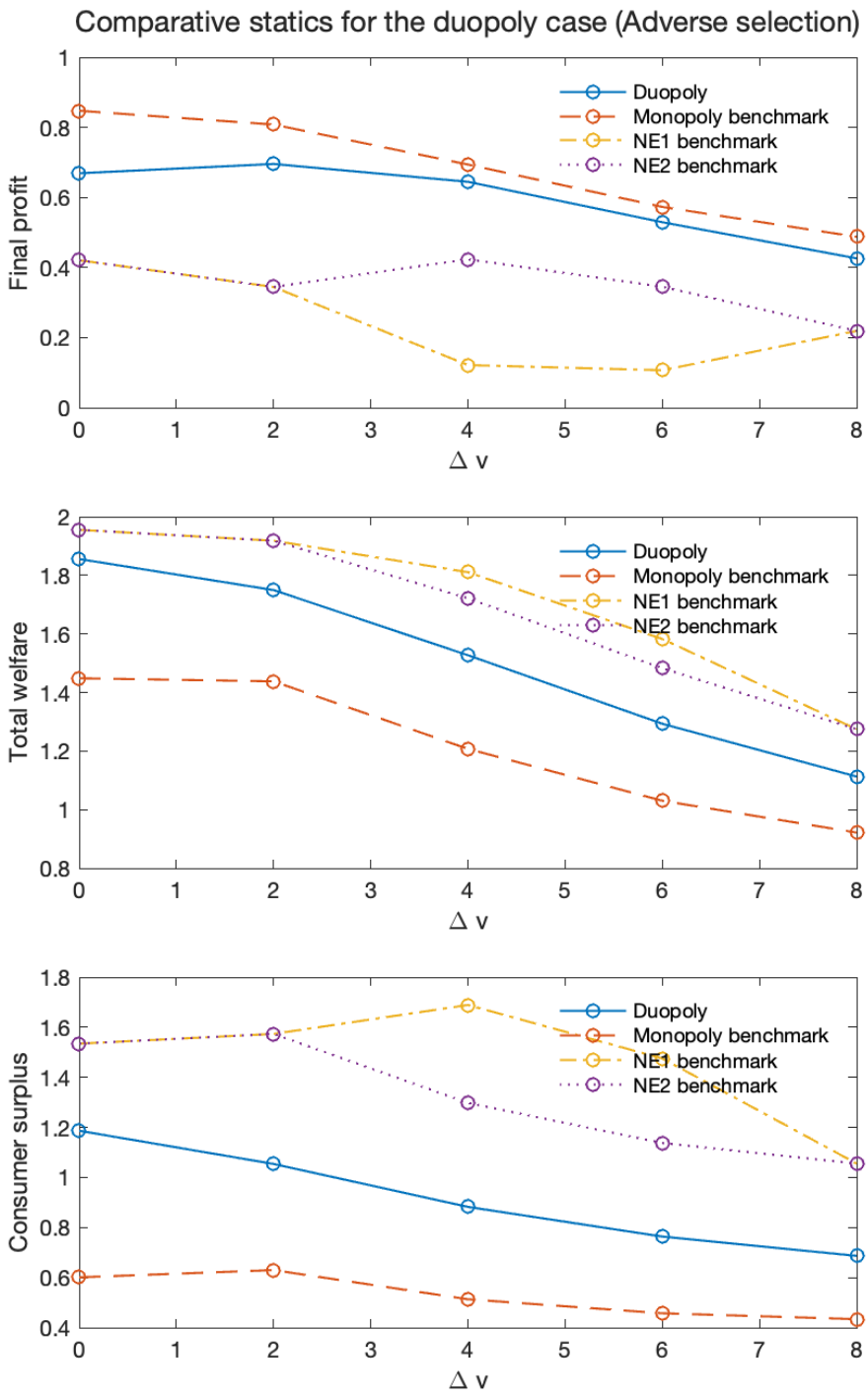


(c) Number of AMs (N)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

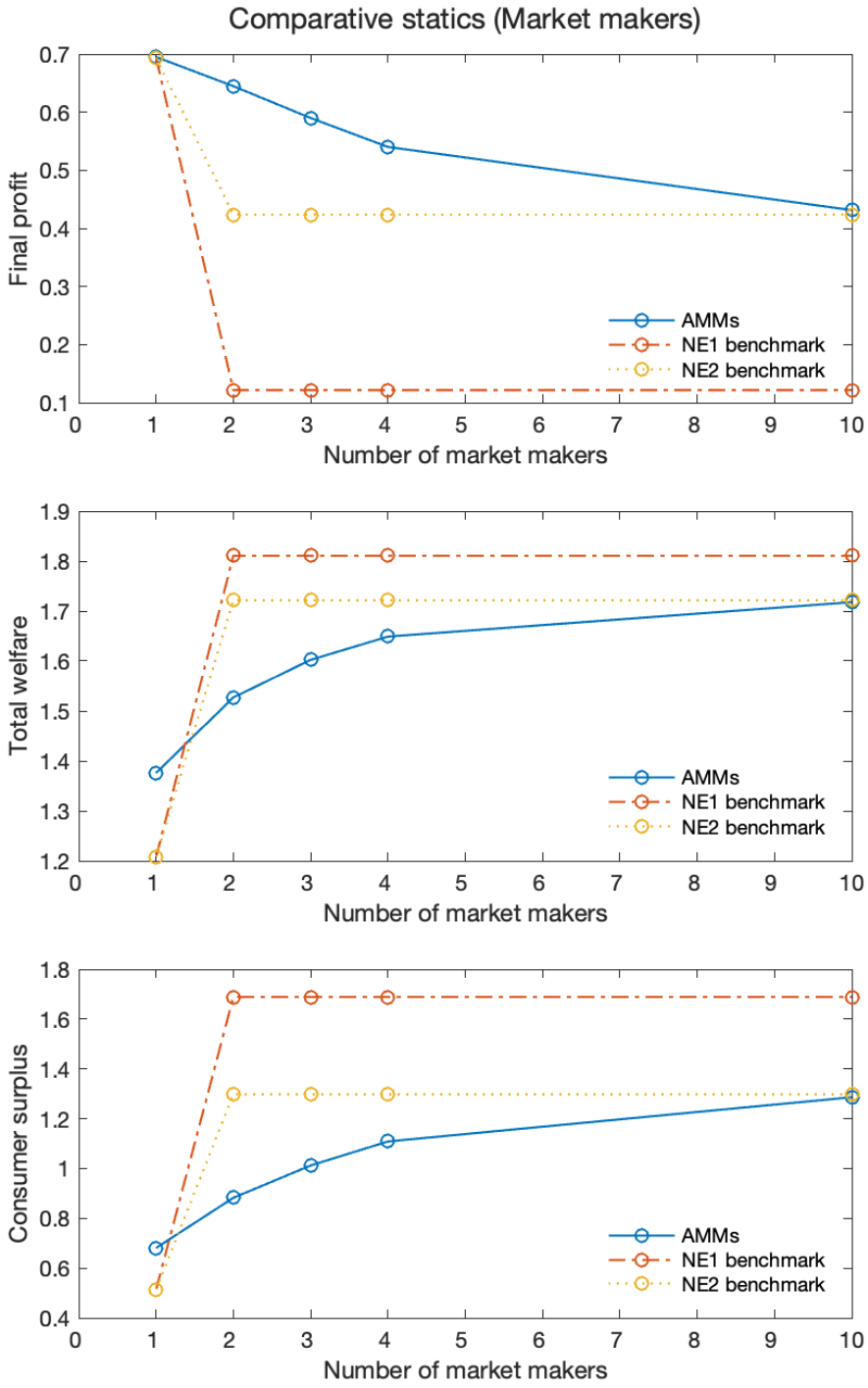
Figure 3: Comparative statics



(a) Dispersion of Clients' Private Valuations (σ)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

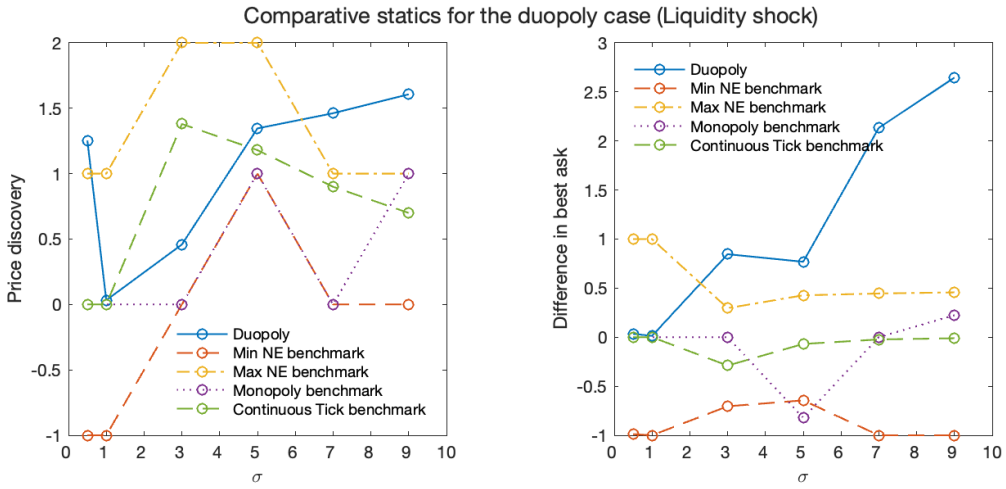


(b) Volatility of the Asset Payoff (Δv)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$

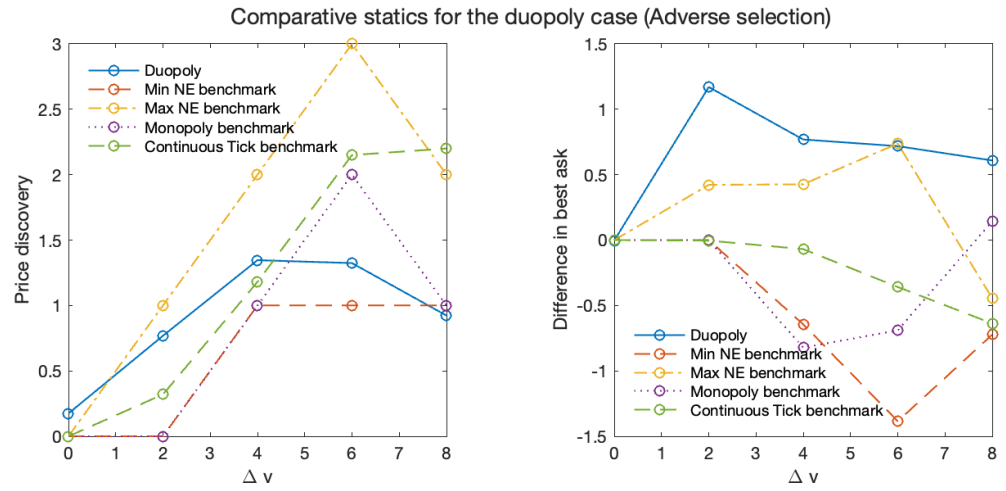


(c) Number of AMs (N)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)

Figure 4: Comparative statics



(a) Dispersion of Clients' Private Valuations (σ)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$ ($v_H = 4$, $v_L = 0$ and $\Delta v = 4$)



(b) Volatility of the Asset Payoff (Δv)
Clients' private valuations are normally distributed with mean zero and variance σ^2. Moreover, $\sigma = 5$, $\mathbb{E}(v) = 2$ and $\mu = \frac{1}{2}$

Figure 5: Comparative statics

A.3 Derivation of the Competitive Price

In this section, we explain how to compute the competitive price in a given trading round for given dealers' beliefs about the distribution of the asset payoff. We do so in the general case (for any $\overline{\tau}$) so that our results apply in particular when $\overline{\tau} = 1$ and $\overline{\tau} = 2$.

Let $V_\tau = I(a(\tau), \tilde{L}_\tau, \tilde{v}) \in \{0, 1\}$ denote the realized trade in period τ and let $I_{n\tau} = V_\tau Z(a_{n\tau}, a(\tau)) \in [0, 1]$ the trade executed by dealer n in round τ. Let H_τ denote the trading history (the observation of clients' trading decisions and the best quotes until trading round τ). That is, $H_\tau = \{(V_i, a_i^{min})\}_{\{i=1,2,...\tau,\}}$ for $\tau \geq 0$ and $H(0) = \emptyset$. The trading history contains information about the asset payoff. Indeed, holding the best quote constant, a client is more likely to buy the asset when $\tilde{v}$ is large than when $\tilde{v}$ is low. Let $\mu(H_{\tau-1})$ be dealers' estimate of the probability that $\tilde{v} = v_H$ at the beginning of trading round τ (with $\mu(0) = \mu = \frac{1}{2}$), given the trading history.

In a Nash-Bertrand equilibrium, in the τ^{th} trading round, all dealers posts the same price a_τ^c such that their expected profit is zero. This happens only if the expected profit of the dealer posting the lowest price is nil among all dealers. Thus, a_τ^c solves:

$$\overline{\Pi}(a_\tau^c, \mu(H_{\tau-1})) = \mu(H_{\tau-1})D(a_\tau^c, v_H)(a_\tau^c - v_H) + (1 - \mu(H_{\tau-1}))D(a_\tau^c, v_L)(a_\tau^c - v_L) = 0. \quad \text{(A.1)}$$

We deduce that:

$$a_\tau^c = \mathbb{E}(\tilde{v} \mid H_{\tau-1}) + \frac{\mu(H_{\tau-1})(1 - \mu(H_{\tau-1}))(v_H - v_L))(D(a_\tau^c, v_H) - D(a_\tau^c, v_L))}{\mu(H_{\tau-1})D(a_\tau^c, v_H) + (1 - \mu(H_{\tau-1}))D(a_\tau^c, v_L)}. \quad \text{(A.2)}$$

The competitive price is the smallest solution to this equation. Observe that it is equal to dealers' expectation of the asset payoff conditional on their information at the beginning of trading round j plus a markup (since $D(a_\tau^c, v_H) - D(a_\tau^c, v_L) = G(^c - v_H) - G(a^c - v_L) > 0$). This markup increases with dealers' uncertainty about the asset payoff at the beginning of trading round τ (measured by $\mu(H_{\tau-1})(1 - \mu(H_{\tau-1}))(v_H - v_L)$).

There is no analytical solution to (A.2). However, one can easily solve it numerically for specific parameter values. To solve (numerically) for the competitive price in the first trading round, we just replace $\mu(H_{\tau-1})$ by $\mu = 1/2$ in (A.2) (dealers' prior at the beginning of an episode). To solve

for the competitive price in the second trading round after a trade in the first trading round, we replace $\mu(H_1)$ by $\mu_2(1, a_1^c)$ (given in (18) in the text) in (A.2). To solve for the competitive price in the second trading round after no trade in the first trading round, we replace $\mu(H_1)$ by $\mu_2(0, a_1^c)$ (given in (19) in the text) in (A.2). Also, note that the probability that a trade occurs in trading round τ is $Pr(V_\tau = 1) = \mu(H_{\tau-1})D(a_\tau^c, v_H) + (1 - \mu(H_{\tau-1}))D(a_\tau^c, v_L)$. Hence, one also gets that:

$$a_\tau^c = \mathbb{E}(v \mid H_{\tau-1}, V_\tau = 1). \tag{A.3}$$

That is, the competitive price is the expected payoff of the asset conditional on the beginning of the trading history up trading round τ and the occurrence of a trade in trading round τ.

A.4 Derivation of the Monopolist's Prices

For any given belief $\mu \in [0, 1]$ that a monopolist might have about $\tilde{v}$ in a given round τ, if the monopolist sets a price of a, his expected payoff from that trading round is equal to $\bar{\Pi}(a, \mu_{\tau-1})$. Let $a^m(\mu)$ solve:

$$a^m(\mu) \in Arg \max_a \bar{\Pi}(a, \mu). \tag{A.4}$$

That is, $a^m(\mu)$ is the price that maximizes the monopolist dealer's expected payoff in round τ, given his belief μ. If the monopolist plays price $a^m(\mu)$, his round τ expected payoff is equal to

$$\Pi^*(\mu) := \bar{\Pi}(a^m(\mu), \mu)$$

Monopolist case with one trading round ($\bar{\tau} = 1$). Because the initial belief is $\mu = \frac{1}{2}$, when there is a single trading round, the monopolist sets a price of $a^m\left(\frac{1}{2}\right)$.

Monopolist case with two trading rounds ($\bar{\tau} = 2$). We now consider the optimal pricing policy of a monopolist dealer when there are two trading rounds. To do so, we proceed by backward induction.

In the second and last round, the price that the monopolist will choose if at the beginning of the second period his belief is μ_2 must be equal to $a^m(\mu_2)$ leading to a second period payoff of $\Pi^*(\mu_2)$

Let consider now the monopolist total payoff from the perspective of period 1. If he sets a first period price of a, then his posterior belief μ_2 is equal to $\mu_2(1,a)$ (given in (18) in the text) in (A.2) if the first period client buys, whereas $\mu_2 = \mu_2(0,a)$ (given in (18) in the text) if the first period client does not buy. Hence the monopolist will set his first period price a_1^m equal to the price a maximizing his total payoff

$$\underbrace{\bar{\Pi}\left(a, \frac{1}{2}\right)}_{\text{First round payoff}} + \underbrace{Pr(a)\Pi^*(\mu_2(1,a)) + (1 - Pr(a))\Pi^*(\mu_2(0,a))}_{\text{second round payoff}}, \tag{A.5}$$

where $Pr(a) := \frac{1}{2}(D(a, v_H) + D(a, v_L))$ is the probability that a trade takes place at date 1 if the monopolist chooses price a at this date. Thus, in choosing her price at date 1, the monopolist accounts for the effect of this price on her expected profit on the trade at date 1 and her continuation value.

When $\bar{\tau} = 2$, we obtain the benchmark price at date 2 in the monopoly case by solving numerically (A.4) (both when there is a trade at date 1 and when there is no trade) and the benchmark price at date 1 by maximizing (A.5).

A.5 Proof of Lemma 1

Fix a price a_m and a dealer n. Suppose that at episode t the dealer's price is $a_{n,t} = a_m$ and it is the lowest price among dealers, i.e. $a_{n,t} = a_m = a_t^{\min}$. Then three outcomes are possible: either the dealer does not trade, the dealer sells the asset worth v_H, or the dealer sells the asset worth v_L. In all cases the Q-matrix is updated. If the dealer does not trade then $\pi_{n,t} = 0$ and $q_{m,n,t+1} = (1-\alpha)q_{m,n,t+1}$, implying

$$|q_{m,n,t} - q_{m,n,t+1}| = \alpha |q_{m,n,t}|$$

If the dealer trades then $q_{m,n,t+1} = \alpha(a_m - \tilde{v}) + (1-\alpha)q_{m,n,t+1}$, and thus

$$|q_{m,n,t} - q_{m,n,t+1}| = \alpha |a_m - v_H - q_{m,n,t}|$$

if $\tilde{v} = v_H$, and

$$|q_{m,n,t} - q_{m,n,t+1}| = \alpha |a_m - v_L - q_{m,n,t}|$$

if $\tilde{v} = v_L$. Denote $\Delta_m(q) := \alpha \max\{|q|, |a_m - v_H - q|, |a_m - v_L - q|\}$ the maximum possible value of that $|q_{m,n,t} - q_{m,n,t+1}|$ can take given that $q_{m,n,t} = q$. Note that

$$\min_q \Delta_m(q) = \alpha \max\left\{\frac{a_m - v_L}{2}, \frac{v_H - a_m}{2}, \frac{v_H - v_L}{2}\right\} = \frac{\alpha}{2}\left(v_H - v_L + \left|a_m - \frac{v_H - v_L}{2}\right|\right) = \Delta_m^*$$

In words, no matter the value of $q_{m,n,t}$, at least one of the three possible outcomes mentioned above leads to $|q_{m,n,t} - q_{m,n,t+1}| \geq \Delta_m^*$. Thus the probability that $|q_{m,n,t} - q_{m,n,t+1}| \geq \Delta_m^*$ cannot be smaller than the smallest of the probabilities of these three events.

Now, given $a_{n,t} = a_m = a_t^{\min}$, the probability that the dealer sells the asset worth v_H, is at least $\frac{1}{2N} D(a_m, v_H)$. The probability that the dealer sells the asset worth v_L, is at least $\frac{1}{2N} D(a_m, v_L) < \frac{1}{2N} D(a_m, v_H)$. The probability that the dealer does not trade is $1 - \frac{1}{2}(D(a_m, v_L) + D(a_m, v_H))$, hence the expression for P_m^*. Q.E.D.

Uncovering the Liquidity Premium in Stock Returns Using Retail Liquidity Provision*

Yashar H. Barardehi Dan Bernhardt
Zhi Da Mitch Warachka

April 9, 2023

Abstract

In response to institutional liquidity demand, wholesalers internalize retail trades. The resulting imbalances in internalized retail order flow coincide with institutional price pressures whose reversals yield a positive relation between these imbalances and future returns. We measure stock-level illiquidity using the likelihood/intensity with which wholesalers facilitate such retail liquidity provision to institutions. Unlike existing illiquidity measures, these easy-to-construct new measures have economically-meaningful relations with institutional holding horizons at stock and investor levels, and yield annualized liquidity premia of 2.7–3.2% post-2010. Thus, we uncover a channel through which a subset of internalized retail order flow predicts the cross-section of returns.

Keywords: Cross-section of Stock Returns, Microstructure, Institutional Trading Costs, Internalized Retail Trade, Liquidity Premium

*We thank Yakov Amihud, James Angel, Azi Ben-Rephael, Hendrik Bessembinder, John Campbell, Amy Edwards, Greg Eaton, Tom Ernst, Daniel Gray, Björn Hagströmer, Terry Hendershott, Paul Irvine, Charles Jones, Alla Kammerdiner, Mete Kilic, Pete Kyle, Marc Lipson, Liang Ma, Albert Menkveld, Dermot Murphy, Shawn O'Donoghue, Michael Pagano, Cameron Pfiffer, John Ritter, Thomas Ruchti, Gideon Saar, Chris Schwarz, Andriy Shkilko, Chester Spatt, Jose Tessada, as well as seminar and conference participants at Cal Poly - SLO, the California Corporate Finance Conference, the Microstructure Exchange, Microstructure Seminars - Asia-Pacific, 2022 Santiago Finance Workshop, 2022 FMA Annual Meetings, and 2023 Finance Down Under for helpful comments. This paper incorporates results from "Internalized Retail Order Imbalances and Institutional Liquidity Demand." Barardehi (barardehi@chapman.edu) is at the Argyros School of Business & Economics, Chapman University and the U.S. Securities and Exchange Commission. Bernhardt (danber@illinois.edu) is at Department of Economics at the University of Illinois and the University of Warwick. Da (zda@nd.edu) is at the Mendoza College of Business, University of Notre Dame. Warachka (warachka@chapman.edu) is at the Argyros School of Business & Economics, Chapman University. Institutional liquidity measures developed by this paper are available at Yashar Barardehi's website. Any errors are our own.

1 Introduction

Various studies have documented that retail order flow predicts the cross-section of stock returns. However, the source of this predictive power is less clear. Some studies simply attribute this return predictability to a subset of retail investors possessing stock-specific information (e.g., Kelley and Tetlock (2013), Fong, Gallagher, and Lee (2014), and Boehmer, Jones, Zhang, and Zhang (2021)). Conversely, Kaniel, Saar, and Titman (2008) argue that retail investors may effectively trade against institutional investors whose trades exert price pressures that eventually reverse, leading to a positive association between retail order flow and future returns.[1] The challenge for testing this mechanism is that order flow segmentation prevents institutional investors from directly interacting with marketable retail order flow. We address this challenge by using microstructure features of modern U.S. equity markets that allow publicly available data to uncover an economic mechanism underlying *indirect* retail-institutional order flow interactions. We establish that absolute imbalances in an easily-observable subset of retail trades provide novel measures of stock liquidity that also capture implicit institutional trading costs. We then document the strong explanatory power of these liquidity measures for expected returns, uncovering a new channel through which retail order flow predicts stock returns.

We provide the first evidence of wholesalers intermediating between retail and institutional investors in modern equity markets, wherein a wholesaler chooses to "internalize" unequal amounts of retail buy vs. sell orders to offset inventory accumulated from providing liquidity to institutional investors on the opposite side of the market. We obtain imbalances in long-only institutional and short-seller trading interests from ANcerno and FINRA data that we link to imbalances in a select subset of internalized marketable retail orders identified using the algorithm proposed by Boehmer et al. (2021), henceforth BJZZ.[2] Crucially, the BJZZ algorithm differentially identifies a subset of retail orders that wholesalers internalize to provide liquidity to institutions.[3]

[1] To clarify, we are interested in *unconditional* return predictability of retail order flow. Some studies examined this return predictability conditional on imminent earnings announcement (e.g., Kaniel, Liu, Saar, and Titman (2012); Boehmer et al. (2021)).

[2] Importantly, using data from 58 brokers and 6 wholesalers, SEC (2022) implies BJZZ's algorithm identifies less than 40% of all marketable retail orders. Barber, Huang, Jorion, Odean, and Schwarz (2022), using self-generated trades, and Battalio, Jennings, Salgam, and Wu (2022), using proprietary wholesaler data, obtain similar conclusions.

[3] Battalio et al. (2022) also find BJZZ's algorithm might mis-classify institutional trades as retail trades. Robustness analyses, reported in Section 5.2 and Internet Appendix C.3, indicate that this does not impact the algorithm's ability to identify retail trades internalized by wholesalers to provide liquidity to institutional clients.

Like BJZZ, we find imbalances in internalized marketable retail flow, denoted $Mroib$, vary across stocks and robustly predict future stock returns for several weeks. However, rather than informed retail trading, we attribute this return predictability to the subsequent unwinding of institutional price pressure, consistent with Kaniel et al. (2008).[4] We provide evidence that large imbalances in these observable internalized retail trades—large $|Mroib|$—reflect the internalized retail orders used by wholesalers to balance their inventories when providing liquidity to institutional investors, especially when liquidity is scarce. This leads us to propose stock-level averages of $|Mroib|$ as liquidity measures. These easy-to-construct liquidity measures proxy for cross-sectional variation in institutional trading costs and, unlike existing liquidity measures, are related to investor holding horizons as predicted by theory. In further contrast, our liquidity measures identify annualized liquidity premia of 2.74–3.20%, associated with one standard deviation reduction in liquidity, post 2010 when existing liquidity measures fail to explain the cross-section of expected stock returns.

Figure 1. Retail Imbalances versus Institutional Imbalances and Price Impacts. This figure plots institutional trade imbalances and institutional-trade price impacts constructed from ANcerno data against imbalances in the volumes of observable internalized retail orders ($Mroibvol$). Each week, stocks are sorted into deciles according to their respective internalized retail order flow imbalance. The averages of institutional trade imbalances and institutional price impacts are then calculated within each decile each week using ANcerno data from 2010–2014. Time-series means of these averages are plotted by $Mroibvol$ decile.

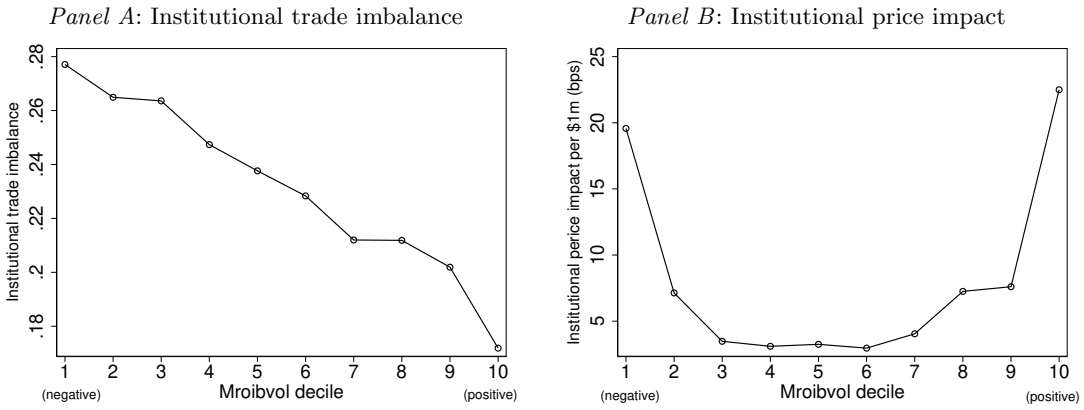


Figure 1 illustrates two properties of $Mroib$ that highlight the liquidity provision facilitated by retail order flow internalization. Panel A shows that institutional trade imbalances are inversely related to BJZZ-identified retail imbalances,[5] while Panel B shows that institutional price impacts

[4] Internet Appendix C.3 proposes improvements to BJZZ's algorithm that reinforce $Mroib$'s return predictability.

[5] Table 2 shows that short sellers are net buyers (sellers) when $Mroibvol$ is negative (positive), even though we have to aggregate observations over bi-weekly horizons rather than daily. The positive average institutional trade imbalance in Figure 1 is expected as mutual funds experienced net inflows in the 2010–2014 post-crisis period.

are highest when these retail imbalances are the most extreme. These patterns suggest that large imbalances in this internalized retail order flow reflect the internalization choices of wholesalers in response to the opposing liquidity demand imbalances of institutions facing high trading costs.

The U.S. equity market structure provides wholesalers, a group of high-frequency market makers, a competitive advantage in providing liquidity to institutions in less liquid markets. Wholesalers interact with institutional investors on exchanges, Alternative Trading Systems (ATSs), and their own Single-Dealer Platforms (SDPs). On the other side, retail brokers outsource the handling of nearly all customer orders to wholesalers in return for payment for order flow (PFOF) or sub-penny price improvements (PI) for their customers. Wholesalers can then choose to (i) internalize retail orders by executing them against their own capital and offering PI; (ii) execute retail orders on a riskless principal basis, without PI, by rerouting orders to ATSs or exchanges; or (iii) reroute retail orders to another wholesaler. Hence, wholesalers secure the option to fill retail orders before these orders are exposed to other market participants, effectively segmenting order flow.[6] Reflecting this segmentation, U.S. wholesalers do not compete with retail investors when providing liquidity to institutions.[7] Instead, wholesalers can use retail flow as an *exclusive* inventory management mechanism (Baldauf, Mollner, and Yueshen (2022)). Thus, wholesalers can offset inventory accumulated from filling unbalanced institutional order flow by *choosing* to internalize disproportionately more retail order flow from the opposing side of the market, especially when liquidity is scarce.[8]

Crucially, the internalized retail orders that facilitate this intermediation often involve sub-penny retail execution prices due to PI and are observable by BJZZ's algorithm. As detailed in Section 3.2, most retail trades not identified by this algorithm are of two types: (1) retail trades chosen by the wholesaler for riskless principal execution on an ATS or exchange; (2) retail orders internalized by the wholesaler due to regulatory requirements rather than by choice.

We document more unbalanced internalized retail flow and higher marginal costs of internalization to wholesalers in the form of greater PI or PFOF when institutional liquidity demand is more

[6] Wholesaler internalization choices determine whether other market participants may directly interact with these retail orders. Practitioners describe internalized orders as "inaccessible liquidity" (Cowen Market Structure 2021).

[7] See Korajczyk and Murphy (2019) for high-frequency market makers' interactions with institutional investors in Canada, where unlike the U.S., all retail orders are routed to public venues, e.g., exchanges.

[8] Simultaneous offsetting of institutional inventory using retail orders also mitigates wholesalers' exposure to toxic (informed) institutional orders. Section 3.2 also notes that wholesalers may use institutional-sourced liquidity to offset inventory accumulated due to retail order flow imbalance. Such executions require abundant liquidity as a wholesaler uses institutional-sourced midpoint liquidity to fill unbalanced retail order flow at the midpoint. Importantly, the BJZZ algorithm facilitates identification of *scarce* liquidity by excluding such mid-point-filled retail trades.

unbalanced and trading costs are higher. This evidence indicates that wholesalers respond to the increased demand for liquidity from liquidity-constrained institutional investors by internalizing costlier retail order flow. Intuitively, wholesalers are willing to exercise their *option* to internalize costlier retail order flow in order to facilitate inventory management when filling unusually profitable institutional orders in less liquid markets. Internet Appendix A provides a simple theoretical framework that links internalization choices to the costs of internalization. We then use exogenous variations in the profits and costs of internalization generated by the Tick Size Pilot to document causally the effect of wholesaler choices on $Mroib$.

Cross-sectional tests highlight the impact of institutional liquidity demand on $Mroib$. Internalization of more (less) retail sell orders than buy orders is associated with higher (lower) net institutional buy volume, and more covering (accumulation) of short interest. Consistent with a lack of institutional counter-parties to offset institutional imbalances on ATSs, a larger $|Mroib|$ is associated with abnormally low quote-midpoint liquidity. In addition, larger $|Mroib|$ is associated with wider quoted spreads and lower quoted depth. These revealed low levels of liquidity present wholesalers with opportunities to fill institutional orders at wide spreads while maintaining a balanced inventory by internalizing costlier retail order flow. Finally, consistent with retail liquidity provision to institutions, but not informed retail trading, contemporaneous *intraday* prices move in the same direction as institutional trade imbalances and thus in the *opposite* direction of $Mroib$.[9]

Cross-sectional regressions of stock returns on $Mroib$ reveal that higher $Mroib$ is associated with higher near-term future weekly returns (through 12 weeks). Consistent with Kaniel et al. (2008), Internet Appendix C attributes the near-term return predictability of $Mroib$ to price reversals following price pressure induced by persistent institutional trading, especially institutional buying (Hendershott and Seasholes (2007), Akepanidtaworn, Di Mascio, Imas, and Schmidt (2020)). Specifically, negative current $Mroib$ (retail selling, institutional buying) is associated with lower future returns for several weeks due to the unwinding of institutional price pressure. Further decomposing daily returns into intraday and overnight returns sheds further light on the liquidity-driven price dynamics, with intraday institutional price pressure being followed by overnight reversals. Crucially, the relation between $Mroib$ and future returns becomes $\cup$-shaped after 6 weeks. This $\cup$-shape pattern persists for well beyond a year, and is consistent with a liquidity premia demanded

[9] $Mroib$ reflects regular-hour trades, making intraday returns the relevant metric.

by institutional investors for holding less liquid stocks, which tend to have high values of $|Mroib|$, and hence give rise to the $\cup$-shape relationship.

The economic mechanism uncovered by our analysis and the availability of data for large cross-sections of stocks motivate our use of $|Mroib|$ to proxy illiquidity and institutional trading costs. We construct stock-level liquidity measures $ILMT$ and $ILMV$ by averaging daily absolute imbalances in, respectively, the number of trades and trading volumes involving BJZZ-identified internalized retail order flow. Comparing these ILMs to existing liquidity measures reveals that they are among the few that are positively related to institutional price impacts in the cross-section.

We then provide direct evidence that ILMs capture the liquidity concerns of institutional investors better than existing measures by linking the liquidity of fund manager holdings based on different liquidity measures to their holding horizon. As Amihud and Mendelson (1986) observe, managers with longer holding horizons should be more willing to invest in illiquid stocks, implying a positive relation between a manager's holding horizon and the measured illiquidity of their equity under management (EUM). We calculate the illiquidity of EUMs using 15 different liquidity measures. For each measure, we examine the relation between the illiquidity of a fund manager's EUM and their holding horizon. Existing liquidity measures all deliver a non-monotone relation between measured EUM illiquidity and holding horizon. In contrast, ILMs induce a more monotone positive relation, consistent with Amihud and Mendelson's prediction.

We then investigate the relation between illiquidity and holding horizon at the stock level. To do this, we calculate the average holding horizon of fund managers in individual stocks (Gaspar, Massa, and Matos (2005); Cella, Ellul, and Giannetti (2013)) and then regress different stock-level liquidity measures in quarter q on the stock's average institutional holding horizon as well as its volatility, market capitalization, and institutional ownership in quarter $q-1$. The R^2s obtained in regressions using ILMs are 3.5-24.2 times larger than those using existing liquidity measures. Moreover, after orthogonalizing ILMs with respect to existing liquidity measures, the residual ILMs continue to exhibit the predicted positive relation with holding horizon. Conversely, the reverse orthogonalizations only deliver the expected relation with holding horizon for quoted spread and quoted depth. In sum, ILMs are the only liquidity measures that have economically meaningful relations with holding horizon at *both* the investor and stock levels.

Next, we establish that ILMs explain expected stock returns. Fama-MacBeth (1973) specifi-

cations regress stock returns in month m on ILMs in month $m-2$ as well as an array of stock characteristic controls.[10] Skipping month $m-1$ ensures that returns in month m are not confounded by short-term reversals following large retail order flow imbalances.[11] As in the prior literature, we find existing high- and low-frequency liquidity measures are not priced (or have negative liquidity "premia") in the 2010–2019 period. In contrast, ILMs are priced with economically significant liquidity premia: a one standard deviation increase in $ILMT$ ($ILMV$) is associated with an annualized liquidity premium of 2.74% (3.20%), comparable to the institutional price impacts computed from ANcerno data that are priced with an annualized premium of 3.8% over 2010-2014.[12]

Portfolio sorts confirm the economic magnitude of the liquidity premia associated with ILMs. Each month, we sort stocks into deciles based on their $ILMT$s or $ILMV$s in month $m-2$, skip month $m-1$, and examine portfolio returns in month m. The high-minus-low return spreads involving deciles 1 and 10, after a Fama-French three-factor adjustment, are 0.86% and 1.06% per *month* for $ILMT$ and $ILMV$, respectively. Value-weighting returns after removing stocks with smallest 20% market-capitalizations, reduces these risk-adjusted returns to 0.58% and 0.46%, respectively. Robustness tests confirm that risk-adjusted return spreads associated with ILMs exceed those based on existing liquidity measures. Moreover, unlike with existing liquidity measures, significant risk-adjusted return spreads are associated with ILMs between intermediate deciles, such as spreads between decile 2 vs. 9, decile 3 vs. 8, and decile 4 vs. 6.

The regression and portfolio results are confirmed by a battery of robustness tests that use alternative estimation approaches, employ specifications that weight observations unequally, and apply various filters that remove small and/or low-priced stocks from the sample. Our highly robust results enable us to conclude that liquidity premia conditional on ILMs hold among stocks that are the most likely to be held by institutional investors. In terms of economic magnitude, a one standard deviation increase in ILMs is associated with annualized liquidity premia between 2.74–3.74%. Similarly, depending on whether "penny stocks" are included in the sample, annualized risk-adjusted return spreads associated with portfolios based on ILMs range between 4.08–15.24%.

Our liquidity measures reveal that stock returns still reflect economically meaningful trading

[10] Internet Appendix H demonstrates robustness to constructing ILMs over three months, $m-4$ to $m-2$.

[11] Consistent with the stock-specific temporal persistence in ILMs, the use of ILMs from month $m-1$ or skipping more than one month leaves our qualitative findings unaffected.

[12] ANcerno data became unavailable in 2015, preventing liquidity premia estimates using institutional price impacts.

costs incurred by institutional investors when entering and exiting stock positions. As reported by Di Maggio, Egan, and Franzoni (2022), institutional price impacts exhibit a standard deviation of 64bps in recent years. This heterogeneity implies investors should demand a liquidity premium that accounts for stock-level institutional price impacts.[13] Our liquidity premia findings are consistent with these trading costs of institutional investors who collectively hold about 70% of publicly-traded equity in the U.S. (Blume and Keim (2012)) in recent years.[14] According to Amihud (2019), "illiquidity has a number of dimensions that are hard to capture in a single measure, including fixed costs, variable costs—price impact costs that increase in the traded quantity—and opportunity costs." The multifaceted nature of liquidity became even more complicated in the post-RegNMS era where spreads are often a few pennies and depth is negligible in fragmented markets. Indeed, a recent literature cautions against using existing liquidity measures to proxy for institutional trading costs post-RegNMS.[15] We overcome the empirical challenges of measuring liquidity in the modern era by developing liquidity measures based on identifiable intermediation by wholesalers between retail and institutional investors when liquidity is scarce. The likelihood and intensity with which wholesalers engage in such intermediation comprise a persistent stock "characteristic" that explains the cross-sectional variation in expected stock returns.

2 Contributions to the Literature

Our paper extends the literature on the relationship between retail order flow and future returns, some of which documents the return predictability of retail order flow.[16] While studies such as Kelley and Tetlock (2013), Fong et al. (2014), and Boehmer et al. (2021) attribute this return predictability to informed retail trades, Kaniel et al. (2008) posit that unbalanced retail order flow

[13] With quarterly re-balancing and a 50% turnover ratio, annualized round-trip execution costs rise by $4 \times 2 \times 0.5 \times 64\text{bps} = 2.56\%$ per year in response a one standard deviation increase in price impacts. This estimate is close to the liquidity premium estimates inferred from our regression analysis, where one standard deviation increase in ILM is associated with 2.47–3.20% increased expected returns.

[14] In contrast, Asparouhova, Bessembinder, and Kalcheva (2010), Ben-Rephael, Kadan, and Wohl (2015), Drienko, Smith, and von Reibnitz (2019), Harris and Amato (2019), and Amihud (2019), among others find vanishing liquidity premia in recent decades using traditional liquidity measures.

[15] Goyenko, Holden, and C. A. Trzcinka (2009), Chordia, R. Roll, and Subrahmanyan (2011), Kim and Murphy (2013), Holden and Jacobsen (2014), Angel, Harris, and Spatt (2011), O'Hara (2015), Eaton, Irvine, and Liu (2021). Barardehi, Bernhardt, and Davies (2019) propose alternative measures.

[16] E.g., Barber and Odean (2000), Barber and Odean (2008), Kumar and Lee (2006), Foucault, Sraer, and Thesmar (2011), Kaniel et al. (2008), Barrot, Kaniel, and Sraer (2016), Kaniel et al. (2012), Kelley and Tetlock (2013), Fong et al. (2014).

reflects strong institutional liquidity demand on the opposite side of the market, which exerts price pressure that subsequently reverses. They suggest institutional investors offer "price concessions" to "entice" retail investors' liquidity provision, a mechanism hard to reconcile with segmented retail and institutional order flows in today's U.S. equity markets. We provide evidence that wholesalers' exclusive access to retail flow allows them to intermediate between retail and institutional investors. These intermediation choices are reflected by the opposite imbalances in internalized marketable retail orders identified using the algorithm proposed by Boehmer et al. (2021), i.e., $Mroib$, especially when liquidity is scarce. Our findings reinforce Barrot et al. (2016)'s notion of unintentional liquidity provision by retail investors; and are consistent Kaniel et al. (2008)'s conclusions in that we find $Mroib$'s return predictability reflects return reversals following institutional investors' consumption of retail-sourced liquidity.[17] Most importantly, we uncover a new channel for return predictability of retail order flow by showing that institutional trading costs and illiquidity can be proxied by $|Mroib|$, which robustly explains the cross-section of expected returns.

We also contribute to a vast literature that designs stock liquidity measures or examines their implications for asset pricing.[18] Our paper develops a proxy of illiquidity using an easily-observable subset of retail trades, distinguishing our liquidity measures from those in the literature. For example, observing the endogenous responses of sophisticated investors to time-varying liquidity, Barardehi et al. (2019) develop trade-time liquidity measures that reflect per-dollar price impacts measured over successive time intervals required for execution of stock-specific fixed dollar values. Bogousslavsky and Collin-Dufresne (2022) use the volatility in total order flow in a given week as a metric of liquidity *risk*, and document its ability to predict next week's return.[19] Finally, we establish the superior performance of our liquidity measures vis à vis sixteen existing liquidity measures along three dimensions: (1) correlation with institutional price impacts; (2) correlation with institutional holding horizons; and (3) robust ability to explain the cross-section of expected returns. Our findings indicate that even though the BJZZ algorithm measures overall retail trading

[17] Theoretical and empirical studies on the link between internalization and market quality includes Battalio and Holden (1995), Battalio, Greene, and Jennings (1997), Battalio, Greene, Hatch, and Jennings (2002), Peterson and Sirri (2003), Parlour and Rajan (2003), Parlour and Rajan (2003), Battalio (2012), and Amirian and Norden (2021).

[18] E.g., Roll (1984), Glosten and Harris (1998), Brennan and Subrahmanyan (1996), Pástor and Stambaugh (2003), Hasbrouck (2009), Goyenko et al. (2009), Chordia et al. (2011), Kim and Murphy (2013), Barardehi et al. (2019), Bogousslavsky and Collin-Dufresne (2022), among many others.

[19] Bogousslavsky and Collin-Dufresne (2022)'s measure is based on second moments, in contrast to most liquidity measures that employ first moments. These authors are interested in identifying *high-frequency* liquidity risk, rather than a persistent stock characteristic that captures the average costs of entering and exiting stock positions.

and order imbalance with large errors, it can be used to construct effective liquidity measures in modern U.S. equity markets.

3 Institutional Details

3.1 Retail Trade Execution

Executions of retail orders in U.S. equity markets are subject to "best execution" principles.[20] Wholesalers, e.g., Virtu and Citadel, handle the vast majority retail orders on behalf of retail brokers, e.g., Charles Schwab and E*Trade. These high-frequency market makers compete over providing execution quality to retail trades (Battalio and Jennings (2022)), ensuring best execution principles are met in addition to providing payment for order flow (PFOF) to certain brokers.[21]

Retail orders handled by wholesalers are executed in two ways. According to SEC (2022) nearly 20% of marketable retail orders are rerouted for riskless principal execution, where a wholesaler quotes an identical order on exchanges/ATSs and fills the retail order once that proprietary order is executed.[22] The remaining 80% of marketable retail order executions are internalized, a process by which wholesalers execute retail order flow against their own inventory.[23] Wholesalers are usually registered brokers, but are not subject to the rules of registered exchanges or ATSs. Most notably, wholesalers can execute trades at sub-penny prices despite the 1¢ minimum tick size. This flexibility allows wholesalers to coordinate with retail brokers and execute retail orders at sub-penny prices reflecting price improvements that fulfill "best execution" duties and improve execution quality.

Panel A in Table 1 reports the distribution of order types across all non-directed orders[24] and all retail volume executed by wholesalers, along with the average PFOF for each order type. Market orders and marketable limit orders account for a disproportionately large share of executed volume receiving PFOF, indicating that wholesalers prefer internalizing marketable orders over

[20] SEC (2021) describes "best execution" as being "at the most favorable terms reasonably available under the circumstances, generally, the best reasonably available price." See FINRA Regulatory Notice 21-23 for more details.

[21] In addition to receiving order flow from brokers, a wholesaler may also receive retail orders from other wholesalers.

[22] Most retail orders originally placed as non-marketable limit orders are routed to exchange limit order books for riskless principal execution. However, a subset of orders organically placed as marketable limit orders become non-marketable when received by the wholesaler due to rapid quote updates.

[23] In May 2012, internalized orders comprised roughly 8% of consolidated volume in NMS stocks (Tuttle (2022)). Reflecting increased retail investor participation, this fraction was 20% in September 2021 (Rosenblatt (2021)).

[24] Retail investors may use a "directed order" to specifying a particular trading venue. However, directed orders comprise a tiny fraction of the orders received by brokers. For example, about 0.01% of the orders received by TD Ameritrade in the first quarter of 2020 were directed.

non-marketable orders. Calculations suggest the share of executed volume of non-marketable limit orders receiving PFOF is only one fourth that of marketable orders. Of note, non-marketable limit orders executed by wholesalers receive over twice as much PFOF per share as marketable orders.

PFOF and PI combine to determine the direct internalization costs to a wholesaler. PFOF and average PI often reflect pre-negotiated terms between brokers and wholesalers, with brokers often trying to obtain the most favorable average PI for their retail customers. However, there is significant variation in PI across individual transactions. Calculations in Section C.3 that compare each execution price with the corresponding NBBO suggest that over 50% of observable internalized marketable orders receive sub-penny PI of no more than 0.1¢. In contrast, underscoring the significant variation in wholesaler internalization costs, over 35% of internalized orders are executed at prices that are inside the NBBO by over 1¢.

Institutional details suggest two channels underlie these large PIs. Most importantly, the Manning rule requires wholesalers with access to proprietary data feeds on odd-lot liquidity to use any inside-quote liquidity to determine best execution terms. Due to the 1¢ tick size, inside-quote odd-lot liquidity is quoted at 1¢ price increments. Thus, when such liquidity exists, to price improve over the "best available price" some internalized marketable retail orders must receive greater-than-1¢ PI. Second, internalized orders executed at prices over 1¢ inside the NBBO may be inside-NBBO non-marketable limit orders, originally placed as marketable orders.[25] Internalizing such non-marketable limit orders is very costly, even when executed at minimal PI because non-marketable orders receive much higher PFOF.

3.2 Implications for BJZZ's Algorithm

Wholesalers internalize about 80% of the marketable retail orders received (SEC (2022)),[26] and BJZZ's algorithm identifies only a select subset of these trades. The algorithm's systematic selection

[25] Consistent with internalization of some non-marketable limit orders, Virtu Financial reports that Virtu "reflects a substantial percentage", but not *all*, of non-marketable orders handled by them on exchanges. That the average PFOF for non-marketable limit orders slightly exceeds 0.3¢ is consistent with competition from exchanges offering such liquidity-making rebates. Spatt (2020) highlights how liquidity fee/rebate tiers incentivize brokers to let wholesalers handle their non-marketable orders because wholesalers receive higher rebates. Upon receipt of a non-marketable order, the wholesaler may execute it on a riskless principal basis by submitting an identically-priced order to an exchange/ATS. If it is executed, the wholesaler fills the standing retail limit order and pays PFOF to the broker.

[26] Wholesalers typically receive four times as much marketable as non-marketable retail order volume, and they internalize a much smaller percentage of those non-marketable orders according to Rule 606 filings, industry reports (Measuring Retail Execution Quality by Virtu Financial), and our analysis of TAQ data.

of a subset of retail trades is *key* to our analysis for at least three reasons.

First, the BJZZ algorithm excludes retail trades filled at the NBBO. Wholesalers have three main options when handling retail orders: (1) internalize them; (2) execute them on a riskless principal basis by rerouting orders to exchanges/ATSs, where non-midpoint sub-penny execution prices are prohibited; and (3) reroute them to another wholesaler. Over 42% (8%) of rerouted (all) retail orders fill at the NBBO (SEC (2022)), implying that the algorithm excludes retail trades that wholesalers *choose* not to internalize.

Second, the algorithm excludes midpoint-filled retail trades that account for a large share of omitted trades and reflect the best execution requirements of brokers. These requirements *force* wholesalers to internalize orders at the midpoint when they detect undisplayed midpoint liquidity, e.g., due to pinging some exchange/ATS for midpoint liquidity. SEC (2022) reports that over 31% of all retail orders are filled at the quote midpoint (also see Battalio et al. (2022)). Importantly, such trades reflect regulatory requirements and not the endogenous internalization choices of wholesalers to source liquidity for their institutional clients. Hence, excluding these trades, which tend to occur when institutional midpoint liquidity is abundant, improves our identification of retail trades internalized by wholesalers to provide liquidity to institutional investors when liquidity is scarce.[27]

Finally, reflecting wholesaler internalization choices, 55% of retail trades reflect non-midpoint internalized orders that receive PI (SEC (2022)), and BJZZ's algorithm picks up such trades with sub-penny PI.[28] Collectively, the BJZZ algorithm, by focusing on a selected subset of retail trades, makes observable those retail trades that wholesalers *choose* to internalize; and this selection underlies the strength of our liquidity measures.

3.3 Wholesalers and Institutional Liquidity Demand

Most wholesalers, including Citadel Securities and Virtu Americas LLC, own Single Dealer Platforms (SDPs). On SDPs, also known as ping pools, a select set of institutions and institutional

[27] Alternatively, midpoint trades may reflect wholesaler competition to provide execution quality (Battalio and Jennings (2022)). Importantly, such executions require abundant liquidity to facilitate wholesaler inventory management, as a wholesaler uses institutional-sourced midpoint liquidity to fill unbalanced retail order flow at the midpoint. Hence, such intermediation should be excluded from an analysis of scarce liquidity, and BJZZ algorithm excludes it.

[28] Less than 1/3 of PI are in round-pennies (SEC (2022)) and not picked up by the algorithm, but such internalized trades likely reflect wholesaler responses to regulatory requirements like the Manning rule when inside quote liquidity exists, indicative of abundant liquidity. SEC (2022) reports that broker-dealers commonly use proprietary order-book data feeds that are more comprehensive than the SIP. Like retail trades filled at the midpoint, the algorithm's exclusion of these trades helps our analysis of wholesaler choices when liquidity is scarce.

brokers trade against the wholesaler.[29] SDPs date back to 2005, and were originally referred to as Electronic Liquidity Providers (BestEx Research (2022)). By 2017, over 2.5% of all trading in NMS stocks occurred on SDPs, comprising roughly 30% of all internalized retail order flow.[30] An institution may "ping" a wholesaler on its affiliated SDP, often using Indication of Interest or Immediate or Cancel orders to signal an unusually high demand for liquidity. This signal encourages the wholesaler to intermediate between retail and institutional investors by providing the institution with liquidity sourced from retail order flow.[31] In 2021, Citadel and Virtu combined to execute almost 17% of consolidated U.S. trading volume by internalizing retail orders, and their affiliated SDPs accounted for over 4% of this volume (Rosenblatt (2021)). Put differently, they internalized about 425 shares of retail orders per 100 shares of institutional orders filled on their SDPs.

When wholesalers use internalized retail buy (sell) order flow to fill unbalanced institutional sell (buy) liquidity demand, the internalized retail orders often receive sub-penny price improvements. Consequently, the corresponding $Mroib$ will be unbalanced and inversely related to institutional liquidity demand. As institutions with high liquidity demand are prepared to pay more to wholesalers, wholesalers can pay higher internalization costs in the form of high PI or high PFOF, internalizing orders that are executed by more that 1¢ inside the NBBO. This leads to a positive relation between $|Mroib|$ and the intensity with which these high-cost retail orders are internalized.

4 Data

To analyze wholesaler intermediation between retail and institutional investors, we construct our sample following BJZZ for the period January, 2010 to December, 2014, covering common shares listed on the NYSE, AMEX, and NASDAQ.[32] We use daily open and close prices from CRSP to calculate daily close-to-close (CC), intraday open-to-close (ID), and overnight, close-to-open (ON) returns. We account for overnight adjustments and, to minimize the impact of bid-ask bounce,

[29] Trading that does not occur on exchanges or ATSs has attracted the attention of regulators. For example, FINRA Regulatory Notice 18-28 describes the nature of SDP trading, a major component of non-ATS trading, and highlights the agency's transparency concerns that led to Regulatory Notice 19-29, which expanded the transparency of OTC trading volume in December 2019.

[30] See Tuttle (2022) and Trader VIP Clubs, 'Ping Pools' Take Dark Trades to New Level, *Bloomberg*, Jan 16, 2018.

[31] For example, VEQ Link, Virtu's SDP, explicitly advertises Virtu's Client Market Making service as the link between its SDP and their retail-broker clients. We emphasize that retail orders are not "redirected" to SDPs. To profit from its intermediation, the wholesaler uses its own capital to fill both institutional orders and retail orders.

[32] We exclude 2015, which is in BJZZ's sample because our ANcerno institutional trade data ends in 2014. Unreported results verify that all findings that do not require ANcerno data are robust to adding 2015.

returns are on based quote midpoints at close. We aggregate daily log-return observations into overlapping 5-day rolling windows to construct daily cross-sections of 5-day (weekly) returns, as in BJZZ. We include observations with a previous-month-end's closing price of at least \$1.

We follow BJZZ to construct measures of observable internalized retail order flow based on the selected sample identified by their algorithm. Using TAQ data, we focus on round-lot off-exchange trades with sub-penny prices.[33] Transactions are classified as retail buy and sell orders if the sub-penny increments exceed 0.6¢ and are below 0.4¢, respectively.[34] We construct daily, normalized measures of imbalance in internalized retail trade frequency and trade volume. $Mroibtrd = (Mrbtrd - Mrstrd)/(Mrbtrd + Mrstrd)$ divides the difference between the number of internalized retail buy and internalized retail sell orders by their sum, while $Mroibvol = (Mrbvol - Mrsvol)/(Mrbvol + Mrsvol)$ is the normalized difference in internalized trade volume. Panel B in Table 1 reports these measures' summary statistics, which closely match those in BJZZ.[35] We then aggregate these daily observations of normalized internalized retail order flow imbalances into overlapping 5-day rolling windows, constructing daily cross-sections of 5-day (weekly) internalized retail order flow imbalances. We also follow BJZZ to construct stock characteristics, including volatility (VOLAT), book-to-market (BM),[36] previous month's return (RET_{-1}), the compound return over the preceding 5 months ($RET_{(-6,-2)}$), and previous month's turnover (TO).

From TAQ data, we match each identified internalized retail transaction with the National Best Bid and Offer prices at the same millisecond. We calculate the daily fractions of internalized retail volume executed at prices that are at least 1¢ better than the NBBO at the time of transaction. We then match 5-day rolling average of these fractions with 5-day (weekly) $Mroib$ measures.

ANcerno data from 2010-2014 provide institutional trade sizes, buy versus sell indicators, execution prices, and stock identifiers. We aggregate institutional buy and sell trades separately at the stock-day level to construct the institutional analogue of $Mroibvol$ denoted $Inroibvol$. To construct institutional price impact measures we calculate volume-weighted average buy and sell execution

[33] As in BJZZ, our findings are robust to including odd-lots.

[34] Internet Appendix C.3 shows that the algorithm mis-classifies subsets of buy and sell orders. Correcting for this mis-classification using quote midpoints marginally reinforces our qualitative findings.

[35] Simple calculations reveal that $Mroib$ daily imbalances are large enough to meet most institutional liquidity demands. The sum $Mrbvol + Mrsvol$ averages over 92k shares, or over \$1.8 million for a \$20 average share price. Hence, a one standard deviation change in $Mroibvol$ is worth over \$800k, which exceeds the \$500k average dollar value of daily institutional trade reported by ANcerno (Hu, Jo, Wang, and Xie (2018)).

[36] Book value is defined as Compustat's shareholder equity value (seq) plus deferred taxes (txdb).

prices across institutional investors for each stock-day. The price impact of a typical institutional buy trade equals the average execution price minus the open price divided by the open price and scaled by the trade's dollar value in millions. Similarly, the price impact of a typical institutional sell trade equals open price minus the average execution price divided by the open price and scaled by the trade's dollar value in millions. We then aggregate institutional trading outcomes over 5-day rolling windows to construct daily cross-sections of 5-day (weekly) institutional trading outcomes.

To analyze liquidity premia, we construct a sample spanning January 2010 through December 2019, of common shares listed on the NYSE, AMEX, and NASDAQ. We construct two daily institutional liquidity proxies as $|Mroibtrd|$ and $|Mroibvol|$. We use WRDS Daily Indicators, TAQ, and CRSP data to construct the following liquidity measures: (1) time-weighted dollar quoted spreads (QSP); (2) time-weighted share depth (ShrDepth); (3) size-weighted dollar effective spread (EFSP); (4) size-weighted dollar realized spread (RESP); (5) size-weighted price impacts (PIMP);[37] (6) monthly estimates of Kyle's λ, constructed by regressing 5-minute returns (calculated from quote midpoints) on the contemporaneous signed square root of net order flow (estimated using the Lee-Ready algorithm) from the respective month;[38] (7) Amvist liquidity measure, defined as the daily ratio of absolute return to turnover; (8) Roll (1984)'s measure of effective spreads; (9) Amihud (2002)'s measure (ILLIQ); (10) Barardehi, Bernhardt, Ruchti, and Weidemier (2021)'s open-to-close measure (ILLIQ_OC); (11 & 12) Barardehi et al. (2019)'s trade-time liquidity measures (BBD and WBBD);[39] (13) our trade-based institutional liquidity measure ($ILMT$), which averages $|Mroibtrd|$; (14) our volume-based institutional measure ($ILMV$), which averages $|Mroibvol|$. We also construct a stock-specific institutional price impact measure (InPrIm) using ANcerno data from 2010–2014 to directly capture post-trade institutional trading costs per \$100k of trade. For each stock-month, we calculate a size-weighted average of institutional price impacts (defined above) associated with individual institutional trades reported by ANcerno.

For all liquidity measures (including $IMLT$ and $IMLV$), we construct two versions; one over a 1-month-horizon that averages daily liquidity proxies and another that averages daily liquidity proxies over rolling three-month windows with monthly updates. For each ILM measure, we also

[37] In unreported analysis, we verify our liquidity measures also outperform spread and price impact measures constructed relative to quote midpoints.

[38] We follow Holden and Jacobsen (2014) in cleaning the data, matching transactions with the corresponding NBBO with millisecond timestamps.

[39] The sample period for these measures is 2010 to 2017 rather than 2010-2019.

calculate corresponding daily averages of the share of volume occurring at sub-penny prices to total daily trading volume. These measures, denoted SPVS, help isolate extreme ILM magnitudes reflecting excessively-infrequent sub-penny trading at the stock level.

We construct a set of stock characteristics for our asset pricing analysis using data from CRSP and Compustat. For stock j in month m, $RET_{j,m-1}$ and $RET_{j,m-2}^{m-12}$, respectively, capture compound returns over the preceding month and the 11 months prior; $M_{j,m-12}$ reflects market-capitalization based on the closing price 12 months earlier; $DYD_{j,m-1}$ reflects dividend yield, i.e., the ratio of total dividend distributions over the 12 months ending in month $m-2$ divided by the closing price at the end of month $m-2$. The book-to-market ratio, $BM_{j,m-1}$, is the most recently reported book value divided by market capitalization at the end of month $m-1$.[40] We obtain three-factor Fama-French betas for each stock from Beta Suite by WRDS. Our approach employs weekly data from rolling horizons that span the preceding 104 weeks, requiring a minimum of 52 weeks. For each stock month, the set of betas represent estimates from the estimation horizon ending in the last week of that month. As in Ang, Hodrick, Zhing, and Zhang (2006), we use a CAPM regression using daily observations in each month to construct monthly idiosyncratic volatility measures.

We construct measures of holding horizon using institutional ownership (13F filings data). Following Gaspar et al. (2005) and Cella et al. (2013), for each institutional investment manager, we calculate a "churn ratio" at the stock-quarter level. For a given manager in quarter q, the churn ratio for an individual stock in her portfolio is defined as the change in the value of that stock in the manager's portfolio relative to that in quarter $q-1$ that is not attributable to variation in its price, divided by the average value of the manager's holdings of that stock in quarters q and $q-1$. We aggregate manager-quarter churn ratios across all managers holding that stock, with each manager's churn ratio weighted by the fraction of institutional ownership held by that manager in the underlying stock. For each stock-quarter, we use the moving average of these weighted mean churn ratios over the preceding four quarters to measure a manager's holding horizon. We also calculate a weighted average churn ratio at the manager-quarter level using each manager's fractional holding in a stock relative to their overall holdings as weights. We define standardized holding horizons at the manager and stock levels using rank statistics of their churn ratios. Specifically, we use 1 minus churn ratio percentile statistics in a quarter to measure institutional holding horizons.

[40] We use the "linktable" from WRDS to match stocks across CRSP and Compustat, dropping stocks without links.

5 Internalized Retail Order Flow Imbalance (*Mroib*)

This section provides cross-sectional evidence of the impact of institutional liquidity demand on $Mroib$. We show that extremely positive or extremely negative $Mroib$ both signify wholesalers intermediating between retail and institutional investors when the demand for liquidity by institutional investors is unbalanced and liquidity is scarce. We then analyze $Mroib$'s return predictability, providing extensive evidence that $Mroib$'s return predictability is not due to informed retail trading but rather the unwinding of institutional price pressure.

5.1 *Mroib* and Trading Activity

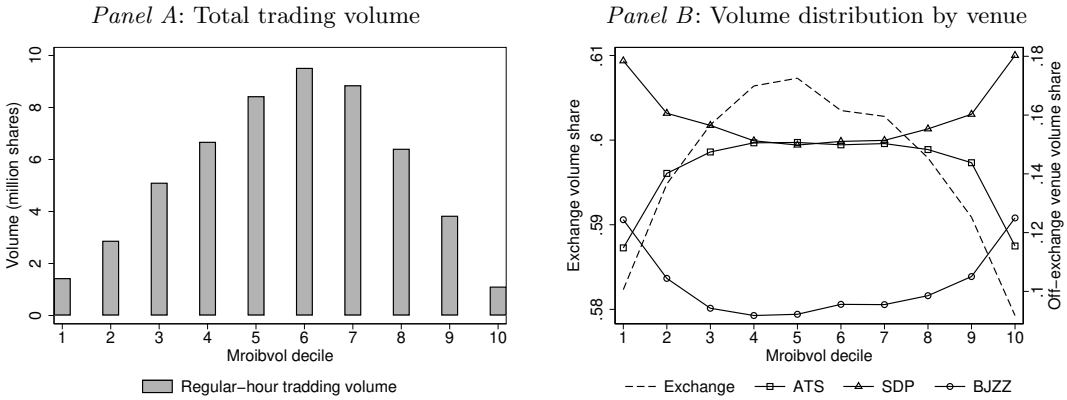
We first examine how $Mroib$ is related to overall trading volume and to the distribution of trading volume across four sources of trading activity: exchanges, ATSs, SDPs, and internalized retail order flow. This analysis provides insights into how each of these sources contributes to the overall trading activity as a function of the prevailing liquidity conditions. We obtain aggregate ATS and non-ATS off-exchange trading volumes at the stock-week level for 01/2019 through 06/2019 from FINRA.[41] Aggregate non-ATS volume is primarily comprised of internalized retail order flow and SDP trading volume (see FINRA Regulatory Notice 18-28). We decompose non-ATS weekly volume into the trading volume identified as retail by BJZZ's algorithm and a residual component. The residual volume is mostly a combination of internalized retail orders executed at (or near) the midpoint and SDP executed institutional volume. Since the midpoint internalized retail volume should be relatively higher when ATS liquidity is high, the opposite should hold for SDP volume.[42] It follows that subtracting this "BJZZ volume" from non-ATS volume yields an over-estimate of SDP volume, especially when ATS midpoint liquidity is high. We construct overall trading volume at the stock-week level from daily observations provided by WRDS Intraday Indicators.

Panel A in Figure 2 reveals that a striking ∩-shaped relationship obtains between trading volume and $Mroibvol$, indicating that a large (absolute) $Mroib$ imbalance is associated with scarce

[41] These data are available from 10/2017. To avoid the effects of the Tick Size Pilot on both ATS and non-ATS volume (Comerton-Forde, Grégoire, and Zhong (2019)), we do not use data from years 2017–2018. Our access to institutional trade data from ANcerno ends in 2014, so we cannot directly examine the relationship between $Mroib$ and institutional trading outcomes, as in Section 5.2, when ATS and non-ATS volume data are available.

[42] Most of the rest of the residual component reflects the internalized retail orders that receive either full-cent PI or zero PI. We show that wholesalers offer greater PI when $Mroib$ is more imbalanced. This suggests that when $Mroib$ is more imbalanced, full-cent PI is more likely and zero cent PI is less likely.

Figure 2. Retail Imbalances versus Trading Volume and Volume Distribution Across Venues. This figure plots total trading volume during regular hours and the cross-venue distribution of trading volume against imbalances in internalized retail order flow ($Mroibvol$). Each *calendar* week, stocks are sorted into deciles according to their respective internalized retail order flow imbalance. Average trading volume as well as average shares of the volume executed on exchanges, on ATSs, on SDPs, and via internalization calculated within each decile each week. Time-series averages of these weekly averages for each decile are plotted from 01/2019 through 06/2019. Weekly ATS and non ATS volumes are obtained from FINRA. The non-ATS volume is decomposed into BJZZ volume, calculated using TAQ data, and SDP volume which is estimated as the difference between non-ATS and BJZZ (internalized retail) volume.



liquidity. Total trading volume is over 80% lower in the extreme (unbalanced) $Mroibvol$ deciles than in the middle deciles that feature near-zero (balanced) $Mroibvol$ levels. The relative absence of trading volume when $Mroibvol$ is unbalanced signifies a decrease in overall liquidity. As a result, the probability that an institutional investor can find another institutional counterparty with whom to trade falls, leaving HFMMs as the primary source for liquidity. Consistent with this, Panel B in Figure 2 presents a break-down of trading volume according to the source of trading activity. The shares of trading volume executed on exchanges and ATS are both over 2bps lower when $Mroibvol$ is at its two most extreme (unbalanced) deciles than when it is close to balanced. The absence of trade on exchanges and ATSs when $Mroibvol$ is most unbalanced is offset by increases of over 2bps in the shares of trading volume executed via SDPs and the internalization of retail order flow. Moreover, (1) BJZZ's algorithm excludes all internalized retail trades executed at the midpoint and (2) midpoint ATS liquidity is notably more abundant when $Mroibvol$ is closer to zero, so our estimates of SDP volume are likely especially biased upwards for intermediate levels of $Mroibvol$. This suggests that the true ∪-shaped pattern of SDP volume share in $Mroibvol$ is *even stronger* than that reported in Figure 2. Importantly, both sources of non-ATS trading activity almost

exclusively reflect wholesaler trades.

These findings indicate that when liquidity is scarce, institutional investors access liquidity provided by wholesalers, encouraging wholesalers to internalize more retail orders. These interactions between institutional investors and wholesalers would lead to imbalances in wholesaler inventory absent their ability to internalize retail orders. To avoid inventory imbalances, wholesalers internalize retail orders, resulting in unbalanced $Mroib$ on the opposite side of the imbalance in institutional order flow. We next provide evidence that $Mroib$ imbalances are in response to wholesalers experiencing a high demand for liquidity from institutions by relating $Mroib$ to institutional order flow imbalances, institutional trading costs and price pressure, and internalization costs.

5.2 *Mroib*, Institutional Trading, and Liquidity

Table 2 summarizes the relationships between $Mroibvol$ and various contemporaneous outcomes across deciles of $Mroibvol$. Close-to-close returns rise monotonically from −2bps in the bottom $Mroibvol$ decile to 30bps in the top decile. However, this pattern is *not* due to price pressure from retail order flow. To show this, we decompose daily returns into intraday and overnight components. Doing so reveals that intraday returns *fall* monotonically from 10bps in the bottom $Mroibvol$ decile to −14bps in the top decile.[43] As most internalized (price-improved) trades are market and marketable-limit orders, the *negative* association between $Mroibvol$ and intraday returns is inconsistent with retail price pressure. This negative association is also at odds with informed retail trading, as it would imply a negative price impact of "informed" orders.

In sharp contrast to intraday returns, overnight returns are positively related to $Mroibvol$. The signs of intraday and overnight returns differ for eight of the ten $Mroibvol$ deciles, in particular for the more extreme, unbalanced $Mroibvol$ deciles. We next investigate different trading outcomes to understand these patterns.

Table 2 shows that, like intraday returns, trade imbalances from both long-only institutional investors and short sellers are negatively related to $Mroibvol$. Average institutional flow falls from 27.7% in the bottom decile to 17.2% in the top decile. Short selling activity also occurs on the opposite side of internalized retail order flow: increased short interest is associated with larger positive internalized retail order flow imbalances. Importantly, directional (as opposed to

[43] Recall that BJZZ's algorithm only uses *regular-hour* off-exchange transactions.

liquidity-providing) short sellers, whose aggregate positions are reflected in short interest data, are known to be informed (Desai, Ramesh, Thiagarajan, and Balachandran (2002); Engelberg, Reed, and Ringgenberg (2012); Boehmer and Wu (2013)). The negative association between such short selling activity and $Mroibvol$ comprises further evidence against the informativeness of retail orders executed at sub-pennies, pointing instead to institutional price pressure driving intraday price movements.

We next show that the negative association between $Mroib$ and institutional trade imbalance does not reflect incorrectly signed institutional trades picked up by BJZZ's algorithm (Battalio et al. (2022)). TAQ data contain ANcerno-reported institutional trades, including those with sub-penny price increments that the algorithm picks up. Battalio et al. (2022) suggest the algorithm incorrectly signs 80% of those trades. To preclude the possibility that $Mroib$ imbalances simply reflect mistakenly-included institutional trade imbalances on the opposite side, we apply the algorithm to execution prices of ANcerno trades to construct BJZZ-implied institutional trade imbalances in ANcerno data. If our results reflect mis-classified institutional trades that enter $Mroib$, then BJZZ-implied institutional trade imbalances must be positively related to $Mroib$. Table 2 shows this imbalance is negative on average, while the analogue for actual institutional imbalance is positive, consistent with Battalio et al. (2022)'s finding that the algorithm signs most institutional trades incorrectly. More importantly BJZZ-implied institutional trade imbalances exhibit no discernible pattern in $Mroib$, establishing that $Mroib$'s negative correlation with ANcerno institutional trade imbalances is a robust feature. Section C.3 provides additional robustness analyses.

We next show that extreme values of $Mroibvol$ are associated with less liquid markets. To do this, we construct a stock-specific measure of abnormal realized off-exchange institutional liquidity. For each stock-day, we divide the volume of large off-exchange mid-point executions[44] by the average of this quantity over the sample period for that stock. Higher values of this measure indicate greater midpoint liquidity. The bottom row in Table 2 shows abnormally low levels of block trades receive off-exchange midpoint execution when $Mroibvol$ is more extreme. That is, large internalized retail order flow imbalances are more common when off-exchange liquidity is abnormally scarce. Together with imbalances in institutional liquidity demand, this finding indicates that institutional investors

[44]TAQ data transactions with trade venue flag 'D' that are at least 1,000 shares, worth at least \$50k, and executed at a price within 0.1¢ of the corresponding quote midpoint.

have trouble locating counter-parties with whom to trade at the midpoint.

Liquidity is also scarce on exchanges when $Mroibvol$ is more extreme. Table 2 shows that spreads are widest and depth at the NBBO is lowest for the extreme deciles of $Mroibvol$. Specifically, median price impacts per \$1m transaction for the average stock are 19bps and 22bps for the lowest and highest $Mroibvol$ deciles, respectively. In contrast, balanced $Mroibvol$ is associated with only 3bps of such costs. Moreover, strikingly, average dollar and relative quoted spreads in the lowest and highest $Mroibvol$ deciles are roughly *double* those when $Mroibvol$ is relatively balanced.

The lack of mid-point liquidity on ATSs means that institutional investors with pressing liquidity needs must turn to venues where they are more likely to trade with HFMMs as intermediaries. Using a wholesaler's SDP allows an institution to trade against a single HFMM—the wholesaler—to conceal its trades. Even when institutional investors opt for exchanges, the exclusive access of wholesalers to segmented retail flow provides them competitive advantages over other HFMMs, making wholesalers more willing to fill institutional orders and thereby creating imbalances in $Mroibvol$.

Importantly, most executions on SDPs and exchanges take place at or near the NBBO because liquidity on these venues is quoted at round-penny increments. In turn, since spreads are wider due to the lack of liquidity, filling institutional demands is unusually lucrative. This suggests that wholesalers may be willing to pay more than normal to internalize retail trade to fill those unusually lucrative institutional orders. Consistent with this argument, the ratio of internalized retail trades executed at prices that are superior to the NBBO by 1¢ or more rises by 33% as $Mroibvol$ diverges from intermediate levels to the two extremes (also see Section C.3). That is, wholesalers incur more costly retail internalization on one side of the market when institutional liquidity demand on the opposite side is abnormally high.

Reverse causality, i.e., wholesalers filling more institutional orders to offset imbalances in internalized retail order flow, cannot explain our findings. For this reverse explanation to hold, the liquidity available to institutions has to *improve* when $Mroib$ is extreme, since wholesalers would need to attract institutional flow by offering abnormally high ATS midpoint liquidity or by improving quoted prices and depth on exchanges. Therefore, under the alternative explanation, an abnormal abundance of retail trading interest on one side of the market would predict that wholesalers internalize retail orders with minimal PI. However, Table 2 reports the exact op-

posite pattern—high $Mroib$ is associated with both higher institutional trading costs and higher internalization costs.[45]

These findings also relate our study to the literature on liquidity timing.[46] Investors with a pressing need to quickly establish or unwind a position may have limited ability to time their trades. This leads Anand, Irvine, Puckett, and Venkataraman (2013) to classify institutional investors as "liquidity demanding" and "liquidity supplying" with the former incurring higher trading costs. Institutional investors accessing liquidity via the internalization of retail order flow in our study are likely "liquidity demanding" institutions. Battalio, Hatch, and Salgam (2022) document higher execution shortfalls for institutional "parent" orders that seek liquidity on SDPs that are typically operated by wholesalers who obtain liquidity by internalizing retail order flow.[47] Our analysis extends these insights by showing that institutions differentially access the liquidity provided by internalized retail order flow when mid-point off-exchange liquidity is scarce. This indicates how wholesalers gain from their access to segmented retail order flow, which they can use for inventory management purposes to offset high institutional demand in less liquid markets.

Our collective findings allow us to attribute the negative association between intraday returns and $Mroib$ to institutional price pressure that occurs in the opposite direction of $Mroib$ imbalances. As such, we reconcile the opposing patterns in overnight returns as price reversals follow institutional price pressure from the preceding intraday period.

Table 2 also reveals that intraday and overnight returns in the extreme $Mroibvol$ deciles reflect more than just the immediate unwinding of price pressure. Most obviously, price pressure from institutional buying is 0.098% in $Mroibvol$'s bottom decile, but the contemporaneous overnight reversal of -0.116% is even larger—a finding that deviates from the stylized fact that unconditional intraday and overnight average returns are negative and positive, respectively (Cliff, Cooper, and Gulen (2008); Berkman, Koch, Tuttle, and Zhang (2012)). To study these phenomena more precisely, we

[45] This is not to say that wholesalers do not use institutional liquidity to provide liquidity to retail investors. Section 3.2 discusses why this type of intermediation, which most likely happens when liquidity is abundant, is not picked up by the BJZZ algorithm, implying that it may not drive our findings.

[46] Research on endogenous liquidity consumption includes Campbell, Ramadorai, and Vuolteenaho (2005), O'Hara (2015), Collin-Dufresne and Fos (2015), Kacperczyk and Pagnotta (2019), and Barardehi and Bernhardt (2021).

[47] This evidence suggests that institutions resort to off-exchange liquidity on SDPs to conceal their intended position sizes by exploiting the delayed reporting of off-exchange trade executions to the Security Information Processor (Ernst, Skobin, and Spatt (2021)). While there may be limited "information leakage" associated with seeking liquidity on SDPs (see BestEx Research), institutional traders have only worse alternatives when mid-point liquidity is limited on ATSs, as trading on exchanges is far more transparent by design.

construct a 5-day overnight return that omits the first close-to-open return and adds the overnight return on the sixth day. This adjustment aligns the timing of intraday price pressure and overnight reversals. This adjustment *exacerbates* the disconnect between the intraday "price pressure" and the subsequent (next-day) overnight "reversals" that average -0.134% when $Mroibvol$ is in decile 1. In fact, comparing intraday and "next-day" overnight returns when $Mroibvol$ is in decile 1 vs. decile 5 reveals differences of $0.098-(-0.063)=0.161\%$ and $-0.0138-0.257=-0.379\%$, respectively. The analogous differences when $Mroibvol$ is in decile 10 vs. decile 5 are $-0.138-(-0.063)=-0.075\%$ and $0.456-0.257=0.199\%$. Thus, weekly overnight returns revert by far more than is needed to offset intraday returns, especially when $Mroibvol$ is extremely negative. Internet Appendix C.1 reconciles this pattern by establishing that institutional buy order flow is more persistent than institutional sell order flow. As a result, institutional buy order flow predicts returns and, in turn, is predicted by retail imbalance (with an inverse relation) over longer horizons. These findings are consistent with Campbell, Ramadorai, and Schwartz (2009).

5.3 Return Predictability of *Mroib*

We next formally examine the return predictability of $Mroib$. Our findings are inconsistent with $Mroib$ capturing informed retail order flow. In contrast, near-term future weekly returns conditional on $Mroib$ are consistent with price reversals following liquidity consumption by institutional investors. We then analyze $Mroibvol$'s long-term return predictability, providing evidence consistent with extreme $Mroibvol$ stocks being less liquid, and hence requiring greater liquidity premia.

Panel B in Table 1 provides summary statistics that closely match those in Table I of BJZZ, confirming that our construction of $Mroibtrd$ and $Mroibvol$ parallels theirs.[48] We estimate the predictability of weekly returns conditional on $Mroibvol$ by estimating:

$$R_{j,w+i}=c_w^0+c_w^1 Mroibvol_{j,w-1}+{c_w^2}^{\top}\text{controls}_{j,w-1}+u_{j,w+i}, \tag{1}$$

where $R_{j,w+i}\in\{CCR_{j,w+i},IDR_{j,w+i},ONR_{j,w+i}\}$ denotes weekly (rolling 5-day) close-to-close, intraday, and overnight returns, respectively, of stock j in week $w+i$. $Mroibvol_{j,w-1}$ denotes the imbalance in the trading volume of internalized retail order flow receiving sub-penny price im-

[48] Slight differences arise since our sample period spans 2010–2014, while BJZZ's spans 2010–2015.

provement in the previous week. We estimate equation (1) to examine $Mroibvol_{j,w-1}$'s return predictability separately for future returns measured over different segments of a day. Control variables include the previous week's return (R_{w-1}) in percentage points, the previous month's return (RET_{-1}), the return over the five months prior to the last month ($RET_{(-7,-2)}$), return volatility (VOLAT), as well as the natural logs of turnover (ln(TO)), market capitalization (ln(Size)), and book-to-market ratio (ln(BM)). As in BJZZ, we estimate equation (1) using Fama-Macbeth regressions, featuring Newey-West corrected standard errors with 6 lags.

Table 3 presents estimation results for week $i = 0$. The second column corresponds to the second column of Table III in BJZZ. Our point estimate ($\hat{c}_w^1$) of 0.087% is nearly identical to their estimate of 0.09%. Coefficients on control variables are also similar to BJZZ's estimates. However, we document a striking difference between $Mroibvol_{w-1}$'s loadings when overnight and intraday returns serve as dependent variables. Specifically, $Mroibvol_{w-1}$ predicts next week's overnight return with the "correct" positive sign, whereas it predicts next week's intraday return with a *negative* coefficient.

These findings are consistent with temporally-persistent institutional price pressures over successive trading sessions and the partial reversals that occur overnight in between daily trading sessions, i.e., overnight. Table 2 established that $Mroibvol$ imbalances were inversely related to both contemporaneous institutional trade imbalance and price pressure, as reflected by intraday returns. Hence the negative predictive power of $Mroibvol$ for future intraday returns is consistent the persistent institutional price pressure across successive trading days. Internet Appendix C.1 provides direct evidence of this using ANcerno data, confirming existing evidence in the literature (e.g., Campbell et al. (2009) and Akepanidtaworn et al. (2020)). The positive association between current $Mroibvol$ and future overnight returns, implies a negative association between current institutional price pressure and future overnight returns. This is consistent with reversals that follow institutional price pressure (Hendershott and Seasholes (2007)). In sum, these findings allow us to attribute $Mroib$'s short-term return predictability to price dynamics driven by institutional liquidity consumption, rather that informed retail trading.

Our analysis of the the link between current $Mroib$ and longer-term future returns reinforces our interpretation that attributes $Mroib$'s short-term return predictability to institutional consumption of retail-sourced liquidity. Kaniel et al. (2008) document stronger such return predictability for less

liquid stocks. Moreover, less liquid stocks are known to command liquidity premia in the form of greater expected returns. Consistent with these insights, Table 4 shows that stocks with more extreme $Mroibvol$ in week $w-1$ are associated with higher returns in the future. Even though week w returns are monotinically positively related to $Mroibvol_{w-1}$, the return difference between the bottom and top deciles of $Mroibvol_{w-1}$ falls rapidly over time, nearly disappearing by week $w+12$. Instead, a striking $\cup$-shaped pattern in close-to-close returns across $Mroibvol_{w-1}$ deciles emerges at week $w+3$, strengthening sharply in subsequent weeks. For example, average week $w+12$'s close-to-close returns in deciles 1 and 10 of $Mroibvol_{w-1}$ (0.15% and 0.18%, respectively) are over double that in decile 6 (0.07%). This $\cup$-shaped pattern holds in all future weeks—future returns are inversely related to negative $Mroibvol_{w-1}$ and positively related to positive $Mroibvol_{w-1}$.[49]

Hence, we relate the $\cup$-shaped pattern in longer future returns to liquidity premia. A liquidity premium associated with expected trading costs as a stock characteristic implies *long-term* return differences according to the level of liquidity. The strong association between liquidity measures, institutional trading costs, and retail order flow internalization suggests that stocks with more extreme $Mroibvol_{w-1}$ are less liquid. Hence, these stocks should command higher *permanent* expected return (higher cross-sectional returns) as compensation that institutional investors require to hold less liquid assets (where entering and exiting positions is costlier), as Amihud and Mendelson (1986) first argued. To make clear that liquidity premia drive the long-term $\cup$-shaped pattern in returns, we focus on lower $Mroibvol_{w-1}$ deciles, where Internet Appendix C.2 provides evidence that the positive relationship between near-term returns and $Mroibvol_{w-1}$ in lower $Mroibvol$ deciles likely reflects extended price reversals following price pressure from previously-accumulated long institutional positions.[50] Clearly, this positive relationship is temporary and is eventually dominated by the liquidity premia that underlie the $\cup$-shaped pattern in longer-term future returns.[51]

[49] See Interned Appendix B for formal estimates of these distinct relationships.

[50] In high $Mroibvol_{w-1}$ deciles, disentangling short-term and long-term effects in close-to-close returns is more difficult since their impacts on returns have the same sign.

[51] Untabulated findings indicate that decomposing close-to-close returns into intraday and overnight components can identify when liquidity premia are realized during the day and contribute to the asset pricing literature documenting time-of-day return disparities that are important to asset pricing anomalies. Our decomposition of close-to-close returns reveals that the $\cup$-shaped pattern in future close-to-close returns as $Mroibvol_{w-1}$ rises from low deciles to high are due to intraday returns. In fact, overnight returns follow a $\cap$-shaped pattern in $Mroibvol_{w-1}$. Thus, we provide an economic mechanism that reconciles why intraday and overnight return anomalies differ—the $\cup$-shaped pattern in intraday returns reflect liquidity premia, and liquidity premia are realized only when there is trade. These findings are complementary to the conclusions of Bogousslavsky (2021), and, contrary to Lou, Polk, and Skouras (2019), provide a rational explanation for the negative correlation between successive intraday and overnight returns.

Investigating $Mroibvol$'s dynamics provides further evidence that $Mroib$ does not reflect informed directional retail trading. Instead, the likelihood and intensity of extreme $Mroib$ occurrences reflect a stock characteristic, indicative of the extent to which institutional investors consume retail-sourced liquidity through wholesalers when liquidity is scarce. This analysis is motivated by BJZZ's finding that $Mroib$ persists over time—their regression of weekly $Mroibvol$ on lagged $Mroibvol$ yields a coefficient of 0.22 (BJZZ, p. 2265). BJZZ use a linear model to estimate the dynamics of $Mroibvol$, but their assumed $AR(1)$ process fails to capture the heterogeneity in the dynamics of retail imbalances. To show this mis-specification we adopt a non-parametric approach to estimate the distribution of $Mroibvol$ in week $w+i$ conditional on week $w-1$.

Panel A in Figure 3 reveals that stock-weeks with extreme negative and extreme positive $Mroibvol$ quantities in week $w-1$ also tend to have extreme imbalances in week $w+12$. This pattern also holds more generally for different weeks $w+i$. Crucially, stocks with extremely negative $Mroibvol$ in week $w-1$ are likely to have extremely negative **or** positive $Mroibvol$ in week $w+12$. Put differently, extreme retail selling "pressure" predicts *both* extreme retail selling *and* extreme retail buying "pressure" 13 weeks forward. So, too, stocks with extremely positive $Mroibvol$ in week $w-1$ are likely to have extremely positive **or** negative $Mroibvol$ in week $w+12$.[52] To show these findings are inconsistent with a linear formulation of $Mroib$'s persistence, we use simulated data from an $AR(1)$ process as a benchmark—Panel B in Figure 3 shows that very different non-parametric estimates obtain from those in Panel A.

Motivated by these collective findings, we next show that $Mroib$ can be used to construct stock liquidity measures that better capture institutional trading costs than existing liquidity measures. Importantly, reflective of their ability to capture liquidity and institutional trading considerations, these measures are strongly priced in the cross-section of stocks, even in recent years.

6 *ILM* Characteristics

This section highlights the important characteristics of our liquidity measures and contrasts them with existing liquidity measures.

[52] Controlling for stock characteristics leaves the qualitative patterns unaffected.

6.1 *ILM*s, Existing Liquidity Measures, and Institutional Price Impacts

To begin, we investigate how institutional liquidity measures (ILMs) are related to key stock characteristics. We then examine how ILMs compare with existing liquidity measures in exhibiting correlations with future post-trade institutional price impacts.

We construct weekly $ILMT$ and $ILMV$ for each stock by averaging $|Mroibtrd|$ and $|Mroibvol|$, respectively, over 5-day rolling windows to obtain weekly observations. We then match these weekly observations with stock characteristics constructed at the end of the preceding calendar month (see Section 4). After excluding stocks whose previous month's closing price are below \$2 (results are robust to excluding stocks with closing prices below \$5), we sort each weekly cross-section into deciles of $ILM \in \{ILMT, ILMV\}$. We then calculate stock characteristic averages by ILM decile and date before computing the time-series averages of these averages across dates by ILM deciles. Table 5 demonstrates that high-ILM stocks, i.e., less liquid stocks according to ILMs, tend to be small growth stocks with relatively poor recent returns and low CAPM betas.

We next show that for less liquid stocks, according to various measures of liquidity, including ILMs, lower liquidity in month $m-2$ is associated with higher realized post-trade institutional price impacts in month m. However, for more liquid stocks, this monotone relationship obtains only based on a handful of liquidity measures, including ILMs. We sort each monthly cross-section in month m into deciles of a given liquidity measure, constructed in $m-2$, with decile 1 (10) containing the most (least) liquid stocks. We then calculate a time-series average of the institutional price impacts of the median stock in each liquidity decile.[53] Panel A in Figure 4 shows that for more liquid stocks (those in liquidity deciles 1–5), future institutional price impacts only rise monotonically with "improved" liquidity as measured by Kyle's lambda, Amihud measures, trade-time liquidity measures, and ILMs—institutional price impacts display no systematic patterns in other liquidity measures. Panel B in Figure 4 shows that for less liquid stocks (liquidity deciles 6–10), worsened liquidity according to most standard liquidity measures (movements from decile 6 to 10) is associated with increased future institutional price impacts. The bottom line is that

[53] Using order statistics rather than simple correlation coefficients lets us identify potential non-linearities and non-monotonicities. Order statistics ensure that the tails of the distributions do not exert undue influence on our estimates and confound interpretations. These considerations are especially relevant for institutional price impacts obtained from ANcerno data that covers less than 7% of CRSP-reported volume for the average stock (3.5% of volume for the median stock). Using stock portfolios rather than individual stocks as test assets sharply reduces measurement error (and noise) that would otherwise impact stock-level estimates.

most liquidity measures can proxy institutional trading costs for less liquid stocks, while a few, including ILMs, also do so for more liquid stocks.[54] The decline in the ability of traditional market microstructure measures to capture these trading costs in the past two decades reflects numerous significant changes to the equity trading environment.

6.2 Persistence of Institutional Liquidity Measures

We next investigate the temporal persistence in $ILMT$ and $ILMV$ at the stock level to determine whether they comprise a stock characteristic. The institutional liquidity measures $ILMT$ and $ILMV$ used in our asset pricing tests average daily $|Mroibtrd|$ and $|Mroibvol|$ observations over one month.[55] To examine the persistence in these measures, we regress $ILMT$ and $ILMV$ on their lags from the six preceding months. These Fama-MacBeth regressions correct for auto-correlated error terms using Newey-West standard errors based on 6 lags, as do the rest of our regression analyses. We exclude stocks priced below \$2, before estimating equally-weighted and value-weighted regressions (with weights computed using a stock's market capitalization in the previous month).

Table 6 documents strong persistence in ILMs: past ILM levels strongly predict future levels. That is, stocks with high ILMs in one month tend to have high ILMs in future months. This holds even when we weight observations by market capitalization, indicating that persistence is not attributable to the illiquidity of small stocks. This persistence indicates that our liquidity measures represent a stock characteristic that is long-lasting enough to impact institutional investors with extended holding horizons and hence justify the existence of a liquidity premium in stock returns.

7 Liquidity and Institutional Holding Horizon

Our next analyses are motivated by the testable hypotheses in Amihud and Mendelson (1986) that (a) at the investor level, investors with longer holding horizons are predicted to hold less liquid stocks, and (b) at the stock level, less liquid stocks are predicted to be held by institutional investors with longer holding horizons.

[54] Internet Appendix D shows that excluding stocks for which sub-penny volume comprises a low share of total volume leavs our qualitative findings unaffected. As such, the prevalence of sub-penny trade execution does underlie the variation in ILM and its ability to proxy institutional trading costs.

[55] Constructions of $Mroibtrd$ and $Mroibvol$ include all transactions. However, our findings are robust to focusing only on round-lot transactions. Odd-lots are only reported by TAQ after 2013.

7.1 Investor-Level Analysis

To calculate the liquidity of an institutional investor's Equity Under Management (EUM), we first calculate the weighted average of each liquidity measure across all stocks held by individual fund managers. We weight observations by the fraction of an investor's total dollar-denominated portfolio value in a stock. Other EUM characteristics, including volatility, market capitalization, and institutional ownership, are computed using a similar methodology in the previous quarter. We follow Gaspar et al. (2005) and Cella et al. (2013) to construct investor-level churn ratios in the previous quarter. The churn ratio captures the frequency at which a fund enters and exits positions, and hence is inversely related to its holding horizon. The churn ratio is calculated at the stock-quarter level, and then weighted by holdings at the manager-quarter level (see Section 4).

We estimate semi-parametric relations at the investor level between EUM liquidity and holding horizons, defined as 1 minus churn ratio percentiles, after controlling for other EUM characteristics. Each quarter, we obtain regression residuals from fitting EUM illiquidity as a function of volatility, market capitalization, and institutional ownership. We then sort each quarterly cross-section into percentile statistics of residual EUM liquidity and holding horizon, independently. Finally, for each liquidity measure, we fit a local polynomial of the residual EUM liquidity percentiles as a function of holding horizon percentile statistics.

Figure 5 illustrates that EUM illiquidity measured by existing liquidity measures, including quoted and relative spreads, quoted depth at best prices, Kyle's lambda, Amihud measure, and trade-time measures display a strong $\cap$-shaped pattern with respect to holding horizon. In contrast, ILM-based EUM illiquidity displays a more monotonically increasing pattern with holding horizon despite flattening for the longest holding horizons.

7.2 Stock-Level Analysis

Institutional investors hold about 70% of U.S. equity, so the relation between holding horizon and liquidity should extend to the individual stock level. That is, less liquid stocks should be held by institutional investors with longer holding horizons after controlling for other stock characteristics.

To test whether different illiquidity measures yield estimates consistent with this prediction, we follow Vovchak (2014). For each stock in each quarter, we first calculate the weighted-average

churn ratio across all investors holding the stock. The weight assigned to an investor's churn ratio is the fraction held by the investor relative to all institutional investment in the stock. We then calculate moving averages over the four preceding quarters for these churn ratios to obtain a stock-quarter measure of institutional turnover. Finally, we regress each liquidity measure at the end of a quarter on the institutional holding horizon percentile (1 minus churn ratio percentile), controlling for volatility, market capitalization, and institutional ownership from the previous quarter. We estimate Fama-MacBeth regressions with Newey-West standard errors based on 6 lags.

Panel A in Table 7 reports that for most liquidity measures, the institutional holding horizon percentile has a coefficient with the expected sign. However, differences show up in R^2 magnitudes. The R^2s associated with $ILMT$ and $ILMV$ are 0.61 and 0.63, respectively, indicating that holding horizon explains a large amount of the variation in investor-level portfolio liquidity based on ILMs. In contrast, the R^2s associated with existing liquidity measures are notably smaller—the next highest R^2 is 0.44 and most are far lower, with some only marginally different from zero.

To further highlight that ILMs better capture the concerns of institutional investors, we orthogonalize the ILM measures with respect to the other liquidity measures. To do this we use Fama-MacBeth regressions, first regressing $ILMT$ and $ILMV$ on existing liquidity measure X, denoting the respective residuals by Z_{ILMT} and Z_{ILMV}. We then examine the ability of holding horizon to explain variation in these residuals. Next, we reverse the specification and regress each existing liquidity measure, separately, on $ILMT$ and $ILMV$, denoting these respective residuals as Y_{ILMT} and Y_{ILMV}. Finally, we examine the ability of holding horizon to explain variation in these residuals.

The top four rows in Panel B of Table 7 report that, relative to every existing liquidity measure, $ILMT$ and $ILMV$ have incremental liquidity-related implications for institutional investors. In contrast, the bottom four rows in Panel B of Table 7 report that the coefficients for holding horizon have their expected sign *only* for dollar quoted/effective spread, relative effective spread, and quoted depth. Moreover, the R^2s in these specifications indicate that for these four liquidity measures, the variation in the Y_{ILMT} and Y_{ILMV} residuals explained by holding horizon (and stock characteristics) is less than one-twentieth of the variation in the Z_{ILMT} and Z_{ILMV} residuals explained by holding horizon (and stock characteristics). That is, institutional holding horizons better explain ILM residuals than they explain residuals of existing liquidity measures. In sum,

ILMs have incremental implications for investors relative to existing liquidity measures, but the converse is not true.

Overall, ILMs are the only liquidity measures whose relations with holding horizons at *both* at the investor and stock levels match the prediction of Amihud and Mendelson (1986).

8 Liquidity Premia

We next contrast the extent to which ILMs and existing liquidity measures predict the cross-section of expected stock returns over the recent 2010–2019 period. We show that, unlike existing measures, ILMs robustly predict the cross-section of stock returns, with economically-large liquidity premia. Long-short portfolios reinforce these findings.

8.1 Regression Analysis

To examine the abilities of ILMs and the other liquidity measures described in Section 4 to predict future monthly returns, we first estimate the following Fama-MacBeth regression

$$RET_{j,m} = \gamma_m^0 + \gamma_m^{LIQ} \left(LIQ_{j,m-2}\right) + \Gamma^\top \text{CONT}_{j,m-1} + u_{j,m}, \tag{2}$$

with Newey-West-corrected standard errors using 6 lags where the dependent variable $RET_{j,m}$ is stock j's return in month m. $LIQ_{j,m-2}$ denotes one of the liquidity measures obtained at the end of month $m-1$ for stock j. $\text{CONT}_{j,m-1}$ denotes a vector of control variables containing betas from the three-factor Fama-French model, book-to-market ratio, market capitalization, dividend yield, idiosyncratic volatility, and the previous month's return as well as the return from the prior 11 months. Green, Hand, and Zhang (2017) examine the return predictability of a comprehensive list of 94 stock characteristics and find their predictive power to fall sharply after 2003. It is therefore unlikely that controlling for more stock characteristics would qualitatively change our results, as our sample starts in 2010. Consistent with this, our findings are robust to using panel regressions that control for unobserved heterogeneities using stock and date fixed effects.

Recall that we impose a \$2 minimum price requirement to preclude the possibility that findings are driven by penny stocks. To further ensure that results are not spurious, we add a one-month

lag between the construction of each liquidity measure and monthly returns.

Panel A in Table 8 reports that unlike other liquidity measures, both the institutional price impact measure (InPrIm) and the ILMs explain the cross-section of expected returns.[56] Specifically, InPrIM, $ILMT$, and $ILMV$ coefficients are 0.029, 1.20 and 1.27, respectively. Multiplying these coefficients by their respective standard deviations (of 0.109, 0.19, and 0.21) yields monthly liquidity premia of 31.6 bps, 22.8bps, and 26.7bps, respectively. Thus, one standard deviation increases in ILMs are associated with 22.8–26.7bps increases in expected monthly returns, with associated annualized increases of 2.74–3.20%. The analogous annual liquidity premium attributable to realized institutional price impacts is 3.8%. These results comprise strong evidence that investors demand economically-significant liquidity premia.

Online Appendix E documents robustness to \$1 and \$5 minimum share price requirements. Consistent with Barardehi et al. (2019) and Barardehi et al. (2021), quoted depth, $ILLIQ\_OC$, BBD, and $WBBD$ only explain the cross-section of stock returns when a \$1 minimum price filter is imposed, indicating that these measures are only priced in very illiquid stocks. Furthermore, consistent with low institutional trading in penny stocks, InPrIM is not priced with a \$1 minimum price filter, but it is priced with a \$5 minimum price filter.[57]

Panel B in Table 8 presents the significant incremental information content of $ILMT$ and $ILMV$ vis à vis each existing liquidity measures. Each ILM measure is first regressed on an alternative liquidity (price impact) measure using Fama-MacBeth regressions. The residual from such regressions are then used, one at a time, as $LIQ_{j,m-2}$ in equation (2). The $ILMT$ and $ILMV$ residuals, with the exception of those orthogonalized to realized institutional price impacts (InPrIm), explain the cross-section of expected returns. Untabulated results verify that the residuals of existing liquidity measures with respect to our measures fail to explain the cross-section of returns.

These results suggest that the literature's conclusion that liquidity premia have disappeared post-decimalization (e.g., Asparouhova et al. (2010); Ben-Rephael et al. (2015)) solely reflect the use of liquidity measures that no longer capture the institutional features of modern equity markets. In particular, tight spreads (often binding at a penny tick) combined with limited depth at the

[56] In unreported results, we compare ILMs to relative (percentage) quoted, effective, and realized spreads, and find ILMs outperform them in all the three dimensions examined.

[57] Online Appendix H establishes the robustness of these results to the construction of our liquidity measures over 3-month rolling windows. This alternative construction results in monthly liquidity premia of 25–31bps, with associated annual liquidity premia of 3.07–3.74%.

NBBO in a fragmented marketplace cannot capture the complicated trade execution strategies institutions adopted in response. In contrast, ILMs are motivated by the actual trading costs of institutional investors and the propensity with which they need to rely on retail-sourced liquidity through wholesalers. As a result, the use of ILMs reveals that institutions account for cross-stock heterogeneity in trading costs when pricing stocks.[58] That $ILMT$ and $ILMV$ do not outperform InPrIm in these residual analyses suggests that only liquidity measures based on proprietary data with limited availability, such as ANcerno data, can possibly compete with ILMs in capturing institutional trading costs.

Table 9 summarizes the results of several robustness tests that confirm the liquidity premia captured by our liquidity measures. First, estimating equation (2) using panel regressions that include date and stock fixed effects and double-cluster standard errors by date and stock leaves our qualitative findings largely unaffected. Second, correcting for market microstructure noise, as in Asparouhova et al. (2010), does not affect the economic significance of the liquidity premia. Third, excluding the smallest 20% of stocks (at the end of the previous month) leaves our qualitative findings unaffected, indicating that the liquidity premia are not a small-stock phenomena. Intuitively, this reflects the relevance of ILMs to institutional investors who tend to hold larger stocks. Fourth, excluding stocks in the bottom 10% of SPVS in each cross-section results in more efficient estimates of liquidity premia. This reflects that ILMs of stocks with low sub-penny volume tend to have higher measurement error. Fifth, weighting observations by stock-level market-capitalizations improves statistical significance of liquidity premia estimates for $ILMT$, but reduces it for $ILMV$. Sixth, excluding the top and bottom 10% of each ILM cross-section increases the precision of liquidity premia estimates and leaves our qualitative findings unaffected. This indicates that estimates are not driven by the tails of the ILM distributions. Indeed, down-weighting (censoring) extreme ILM observations strengthens our results. All robustness tests are implemented separately after imposing minimum share price requirements of \$1, \$2, and \$5. Seventh, we document robustness of liquidity premia across listing exchanges. This final robustness test is motivated by Asparouhova et al. (2010) and Ben-Rephael et al. (2015), who detect liquidity premia post decimalization for NASDAQ-listed firms, but not NYSE-listed firms. Online Appendix H confirms the robustness of

[58] Kyle's λ fails to explain the cross-section of expected returns. This suggests that the conclusions of Huh (2014) that Kyle's λ explained the cross-section of returns in the 1983–2009 period do not extend past 2010.

the liquidity premia when liquidity measures are constructed over 3-month rolling windows.

Overall, our empirical results provide compelling evidence that ILMs predict expected stock returns and are associated with economically significant liquidity premia.

8.2 Portfolio Sorts

This section reports that long-short portfolios based on ILM generate abnormal (risk-adjusted) monthly returns. For each monthly cross-section, we form 10 liquidity portfolios using $ILMT$ and separately using $ILMV$. These portfolios are first formed by sorting the cross-section of stocks into deciles based on the entire CRSP common-share universe before calculating equally-weighted portfolio returns. In additional robustness tests, we form portfolios breakpoints using ILMs of NYSE-listed stocks after removing stocks whose market capitalization is in the bottom 20% before calculating value-weighted portfolio returns.[59] Portfolio returns are calculated as the average return of the stocks assigned to the respective portfolio net of the contemporaneous 1-month Treasury-bill rate. The monthly long-short portfolio return equals the return difference between the least liquid and the most liquid portfolios. Finally, we regress the time-series of individual portfolio returns as well as the time-series of the long-short returns on the Fama-French three factors. The intercept of each time-series regression is the relevant risk-adjusted return (spread), whose significance is assessed using Newey-West standard errors with 6 lags. We apply three different minimum share price filters that remove stocks whose month-end closing price in the prior month is below $p_{min} \in \{\$1, \$2, \$5\}$.

Table 10 reports significant risk-adjusted return spreads between the least liquid portfolio and the most liquid portfolio according to both $ILMT$ and $ILMV$. The portfolio risk-adjusted returns display roughly monotonic patterns, increasing from the most liquid portfolio to the least liquid portfolio. The corresponding return spreads are economically significant, ranging between 0.93% and 1.20% per month in our main sample (Panel B in Table 10) and between 0.41% and 1.27% per month across all specifications. Online Appendix H establishes the robustness of findings to constructing ILMs over 3-month rolling windows, uncovering three-factor return spreads that range from 0.34% to 1.18% per month. Overall, our estimates imply that annualized portfolio return spreads based on ILM range between 4.08–15.24%, with the larger estimates attributable

[59] Conclusions are robust to alternative combinations of break-points, weights, and small-firm filters.

to samples involving involve small, low-priced stocks.

An investigation of ANcerno data suggests that our liquidity premium estimates are plausible manifestations of expected implicit trading costs. Figure 5 suggests a 20bp difference in expected institutional price impacts between stocks in the top and bottom ILMs deciles for a \$2 price filter. Our institutional price impacts estimates (InPrIm) are re-scaled to reflect costs per \$100k of institutional trade size—hence, the 20bp difference can be re-scaled to reflect the variation associated with alternative benchmark trade sizes. To match the 40-120pbs liquidity premia estimates in Table 10, true dollar values for monthly institutional trade volumes in a typical stock should be about \$200-600k , scaling up the benchmark trade size used in our estimates by factors of 2–6. ANcerno data suggest that these benchmarks are reasonable. The median and average dollar value of institutional trades per month in 2010 are about \$110k and \$1,200k, respectively, when we use a \$2 price filter. These values understate true institutional monthly trade volumes because larger institutional investors employ "in-house" trade execution algorithms and do not use Abel Noser's execution quality assessment services—so their trades are not reflected in ANcerno data.

Internet Appendix F repeats the portfolio sorting exercise for alternative liquidity measures using the three minimum price filters. It confirms that ILMs are the only measures for which the long-short portfolio risk-adjusted return spreads reflect liquidity premia close to 1% or higher.

We also find that alphas associated with ILMs survive double sorts that control for key stock characteristics. Internet Appendix G forms an array of 5×5 portfolios that first condition on a stock characteristic (one of market beta, market capitalization, book-to-market ratios, past returns, and the share of sub-penny volume), and then on an ILM. We document liquidity premia for high- and low-beta, small and large, growth and value stocks, past losers and past winners, and stocks with low and high sub-penny executed volume. We then investigate whether trading costs can explain the returns of anomalies based on stock characteristics by switching the order of the double sorts. Consistent with the existing literature (e.g., Lesmond, Schill, and Zhou (2004); Korajczyk and Sadka (2004)), we find that momentum profits do not survive institutional trading costs.

9 Conclusion

Our paper attributes the strong return predictability of the imbalance in retail buy vs. sell orders internalized at sub-penny prices (Boehmer et al. (2021)) to liquidity provision by retail investors to institutional investors (Kaniel et al. (2008)). Importantly, order flow segmentation in U.S. equity markets prevents retail liquidity provision through direct interactions between retail and institutional order flows. We provide the first evidence of wholesalers, a group of high-frequency market makers, intermediating between retail and institutional investors. Wholesalers' exclusive access to internalized retail orders equips them with a competitive advantage in providing liquidity to institutional investors when liquidity is scarce. When liquidity-constrained institutions access liquidity by interacting with a wholesaler on one side of the market, the wholesaler internalizes unequal amounts of retail buy and sell order flow to offset the inventory they would otherwise accumulate when providing liquidity to institutions. We show that such institutional liquidity consumption when liquidity is scarce is associated with institutional price pressure. The subsequent price reversals create a positive association between imbalances in a *select subset* of internalized retail flow that reflect wholesaler choices and future returns. Hence, this return predictability should not be attributed to informed retail trading.

These findings motivate our use of the absolute value of the imbalance in observable internalized retail flow as a stock-level proxy of institutional trading costs—higher such imbalances signify scarce liquidity from the perspectives of institutional investors. We show that, relative to existing measures, our stock-level institutional liquidity measures are more closely linked with realized institutional trading costs and institutional holding horizons. We also provide robust evidence that our liquidity measures are priced in the cross-section of stock returns and yield economically significant liquidity premia post 2010, when existing liquidity measures are no longer priced. This finding is important for three reasons: (1) consistent with nontrivial institutional trading costs, it shows that stock returns still contain liquidity premia, indicating that a recent literature did not find significant liquidity premia only because their measures no longer capture relative trading costs; (2) it uncovers a new channel for return predictability of retail order flow; and (3) it provides researchers with an easy-to-construct measure of stock liquidity that captures the institutional details of modern U.S. equity markets.

References

Akepanidtaworn, K., R. Di Mascio, A. Imas, and L. Schmidt (2020). Selling fast and buying slow: Heuristics and trading performance of institutional investors. *Journal of Finance*. Forthcoming.

Albuquerque, R., S. Song, and C. Yao (2020). The price effects of liquidity shocks: A study of the sec's tick size experiment. *Journal of Financial Economics 138*, 700–724.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets 5*, 31–56.

Amihud, Y. (2019). Illiquidity and stock returns: A revisit. *Critical Finance Review 8*, 203–221.

Amihud, Y. and H. Mendelson (1980). Market-making with inventory. *Journal of Financial Economics 8*, 31–53.

Amihud, Y. and H. Mendelson (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics 17*, 223–2449.

Amirian, F. and L. L. Norden (2021). High-frequency traders and single-dealer platforms. Working Paper.

Anand, A., P. Irvine, A. Puckett, and A. Venkataraman (2013). Institutional trading and stock resiliency: Evidence from the 2007–2009 financial crisis. *Journal of Financial Economics 108*, 773–797.

Ang, A., R. Hodrick, Y. Zhing, and x. Zhang (2006). The cross-section of volatility and expected returns. *Journal of Finance 61*, 259–299.

Angel, J., L. Harris, and C. Spatt (2011). Equity trading in the 21st century. *Quarterly Journal of Finance 1*, 1–53.

Asparouhova, E., H. Bessembinder, and I. Kalcheva (2010). Liquidity biases in asset pricing tests. *Journal of Financial Economics 96*, 215–237.

Baldauf, M., J. Mollner, and B. Z. Yueshen (2022). Siphoned apart: A portfolio perspective on order flow fragmentation. Working Paper.

Barardehi, Y. and D. Bernhardt (2021). Uncovering the impacts of endogenous liquidity consumption in intraday trading patterns. Working Paper.

Barardehi, Y. H., D. Bernhardt, and R. J. Davies (2019). Trade-time measures of liquidity. *Review of Financial Studies 32*, 126–179.

Barardehi, Y. H., D. Bernhardt, T. G. Ruchti, and M. Weidemier (2021). The night and day of Amihud's (2002) liquidity measure. *Review of Asset Pricing Studies 11*, 269–308.

Barber, B. M., x. Huang, P. Jorion, T. Odean, and C. Schwarz (2022). A sub(penny) for your thoughts: Improving the identification of retail investors in TAQ. Working Paper.

Barber, B. M. and T. Odean (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance 55*, 773–806.

Barber, B. M. and T. Odean (2008). The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies 21*, 785–818.

Barclay, M. J. and T. Hendershott (2003). Price discovery and trading after hours. *Review of Financial Studies 16*, 1041–1073.

Barnea, A. and D. E. Logue (1975). The effect of risk on the market-maker's spread. *Financial Analysts Journal 31*, 45–49.

Barrot, J., R. Kaniel, and D. A. Sraer (2016). Are retail traders compensated for providing liquidity? *Journal of Financial Economics 120*, 146–168.

Battalio, R. (2012). Third market broker-dealers: Cost competitors or cream skimmers? *Journal of Finance 52*, 341–352.

Battalio, R., J. Greene, B. Hatch, and R. Jennings (2002). Does the limit order routing decision matter? *Review of Financial Studies 15*, 159–194.

Battalio, R., J. Greene, and R. Jennings (1997). Do competing specialists and preferencing dealers affect market quality? *Review of Financial Studies 10*, 969–993.

Battalio, R., B. Hatch, and M. Salgam (2022). The cost of exposing large institutional orders to electronic liquidity providers. Working Paper.

Battalio, R. and C. W. Holden (1995). A simple model of payment for order flow, internalization, and total trading cost. *Journal of Financial Markets 4*, 33–71.

Battalio, R. and R. Jennings (2022). Why do brokers who do not charge payment for order flow route marketable orders to wholesalers? Working Paper.

Battalio, R., R. Jennings, M. Salgam, and J. Wu (2022). Identifying market maker trades as "retail" from TAQ: No shortage of false negatives and false positives. Working Paper.

Ben-Rephael, A., O. Kadan, and A. Wohl (2015). The diminishing liquidity premium. *Journal of Financial and Quantitative Analysis 50*, 197–229.

Berkman, H., P. D. Koch, L. Tuttle, and Y. J. Zhang (2012). Paying attention: Overnight returns and the hidden cost of buying at the open. *Journal of Financial and Quantitative Analysis 47*, 715–741.

BestEx Research (2022). Accessing single dealer platforms (SDPs) in execution algorithms: Penny-wise and pound-foolish? *White Paper*.

Blume, M. E. and B. D. Keim (2012). Institutional investors and stock market liquidity: Trends and relationships. Working Paper.

Boehmer, E., C. M. Jones, X. Zhang, and X. Zhang (2021). Tracking retail investor activity. *Journal of Finance 76*, 2249–2305.

Boehmer, E. and J. Wu (2013). Short selling and the price discovery process. *Review of Financial Studies 26*, 287–322.

Bogousslavsky, V. (2021). The cross-section of intraday and overnight returns. *Journal of Financial Economics 141*, 172–194.

Bogousslavsky, V. and P. Collin-Dufresne (2022). Liquidity, volume, and order imbalance volatility. *Journal of Finance*. Forthcoming.

Brennan, M. J. and A. Subrahmanyan (1996). Market microstructure and asset pricing: On the compensation for illiquidity in stock returns. *Journal of Financial Economics 41*, 441–464.

Campbell, J. Y., S. J. Grossman, and J. Wang (1993). Trading volume and serial correlation in stock returns. *Quarterly Journal of Economics 108*, 905–939.

Campbell, J. Y., T. Ramadorai, and A. Schwartz (2009). Caught on tape: Institutional trading, stock returns, and earnings announcements. *Journal of Financial Economics 92*, 66–91.

Campbell, J. Y., T. Ramadorai, and T. O. Vuolteenaho (2005). Caught on tape: Institutional order flow and stock returns. Working paper.

Cella, C., A. Ellul, and M. Giannetti (2013). Investors' horizons and the amplification of market shocks. *Review of Financial Studies 26*, 1067–1648.

Chordia, T., R. R. Roll, and A. Subrahmanyan (2011). Recent trends in trading activity and market quality. *Journal of Financial Economics 101*, 243–263.

Cliff, M., M. J. Cooper, and H. Gulen (2008). Return difference between trading and non-trading hours: Like night and day. Working Paper.

Collin-Dufresne, P. and V. Fos (2015). Do prices reveal the presence of informed trading? *Journal of Finance 70*, 1555–1582.

Comerton-Forde, C., V. Grégoire, and Z. Zhong (2019). Inverted fee structures, tick size, and market quality. *Journal of Financial Economics 134*(1), 141–164.

Desai, H., K. Ramesh, S. Thiagarajan, and B. Balachandran (2002). An investigation of the informational role of short interest in the NASDAQ market. *Journal of Finance 57*, 2263–2287.

Di Maggio, M., M. Egan, and F. Franzoni (2022). The value of intermediation in the stock market. *Journal of Financial Economics 158*, 208–233.

Drienko, J., T. Smith, and A. von Reibnitz (2019). A review of the return-illiquidity relationship. *Working Paper*.

Eaton, G. W., R. J. Irvine, and T. Liu (2021). Measuring institutional trading costs and the implications for finance research: The case of tick size reductions. *Journal of Financial Economics 139*, 823–851.

Engelberg, J., A. Reed, and M. Ringgenberg (2012). How are shorts informed? Short sellers, news, and information processing. *Journal of Financial Economics 105*, 260–278.

Ernst, T., J. Skobin, and C. Spatt (2021). The value of off-exchange data. Working Paper.

Fong, K. Y. L., D. R. Gallagher, and A. D. Lee (2014). Individual investors and broker types. *Journal of Financial and Quantitative Analysis 42*, 431–451.

Foucault, T., D. Sraer, and D. Thesmar (2011). Individual investors and volatility. *Journal of Finance 6*, 1369–1406.

Gaspar, J. M., M. Massa, and P. Matos (2005). Shareholder investment horizons and the market for corporate control. *Journal of Financial Economics 76*, 135–165.

Glosten, L. and L. Harris (1998). Estimating the components of the bid-ask spread. *Journal of Financial Economics 21*, 123–142.

Goyenko, R. Y., C. W. Holden, and C. A. C. A. Trzcinka (2009). Do liquidity measures measure liquidity? *Journal of Financial Economics 92*, 153–181.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average U.S. monthly stock returns. *Review of Financial Studies 30*, 4389–4436.

Griffith, T., B. Roseman, and D. Shang (2020). The effects of an increase in equity tick size on stock and option transaction costs. *Journal of Banking & Finance 114*, 1057–1082.

Grossman, S. J. and M. H. Miller (1988). Liquidity and market structure. *Journal of Finance 43*(3), 617–633.

Harris, L. and A. Amato (2019). Illiquidity and stock returns: Cross-section and time-series effects: A replication. *Critical Finance Review 8*, 173–202.

Hasbrouck, J. (2009). Trading costs and returns for U.S. equities: Estimating effective costs from daily data. *Journal of Finance 64*, 1445–1477.

Hendershott, T., A. Menkveld, R. Praz, and M. S. Seasholes (2022). Asset price dynamics with limited attention. *The Review of Financial Studies 35*, 962–1008.

Hendershott, T. and M. S. Seasholes (2007). Market maker inventories and stock prices. *American Economic Review 97*, 210–214.

Ho, T. and H. R. Stoll (1982). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics 9*, 47–73.

Holden, C. W. and S. E. Jacobsen (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance 69*, 1747–1785.

Hu, E. and D. Murphy (2022). Competition for retail order flow and market quality. *Working Paper*.

Hu, G., K. Jo, Y. Wang, and J. Xie (2018). Institutional trading and abel noser data. *Journal of Corporate Finance 69*, 1747–1785.

Jiang, C., T. Likitapiwat, and T. T. McInish (2012). Information content of earnings announcements: Evidence from after-hours trading. *Journal of Financial and Quantitative Analysis 47*, 1303–1330.

Kacperczyk, M. and E. Pagnotta (2019). Chasing private information. *Review of Financial Studies 32*, 4997–5047.

Kaniel, R., S. Liu, G. Saar, and S. Titman (2012). Individual investor trading and return patterns around earnings announcements. *Journal of Finance 67*, 639–680.

Kaniel, R., G. Saar, and S. Titman (2008). Individual investor sentiment and stock returns. *Journal of Finance 63*, 273–310.

Kelley, E. K. and P. C. Tetlock (2013). How wise are crowds? Insights from retail orders and stock returns. *Journal of Finance 68*, 1229–1265.

Kim, S. and D. Murphy (2013). The impact of high-frequency trading on stock market liquidity measures. *Working Paper*.

Korajczyk, R. and D. Murphy (2019). High-frequency market making to large institutional trades. *The Review of Financial Studies 32*, 1034–1067.

Korajczyk, R. A. and R. Sadka (2004). Are momentum profits robust to trading costs? *Journal of Finance 59*, 1039–1082.

Kumar, A. and C. Lee (2006). Retail investor sentiment and return comovements. *Journal of Finance 61*, 2451–2486.

Lesmond, D. A., M. J. Schill, and C. Zhou (2004). The illusory nature of momentum profits. *Journal of Financial Economics 71*, 349–380.

Lou, D., C. Polk, and S. Skouras (2019). A tug of war: Overnight versus intraday expected returns. *Journal of Financial Economics 134*, 192–213.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics 116*, 257–270.

Parlour, C. and U. Rajan (2003). Payment for order flow. *Journal of Financial Economics 68*, 379–411.

Pástor, L. and R. F. Stambaugh (2003). Liquidity risk and price discovery. *Journal of Political Economy 111*, 642–685.

Peterson, M. A. and E. Sirri (2003). Order preferencing and market quality on U.S. equity exchanges. *The Review of Financial Studies 16*, 385–415.

Rindi, B. and I. M. Werner (2019). U.S. tick size pilot. Working Paper.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance 39*, 1127–1139.

Rosenblatt (2021). Trading talk - market structure analysis: A closer look at off-exchange and retail market share.

SEC (2021). Staff report on equity and options market structure conditions in early 2021.

SEC (2022). Proposed rule: Order competition rule. Release No. 34-96495; File No. S7-31-22.

Smidt, S. (1971). Which road to an efficient stock market free competition or regulated monopoly? *Financial Analysts Journal 27*, 18–20.

Spatt, C. (2020). Is equity market exchange structure anti-competitive? Working Paper.

Stoll, H. R. (1976). Dealer inventory behavior an empirical investigation of nasdaq stocks. *Journal of Financial and Quantitative Analysts*, 356–380.

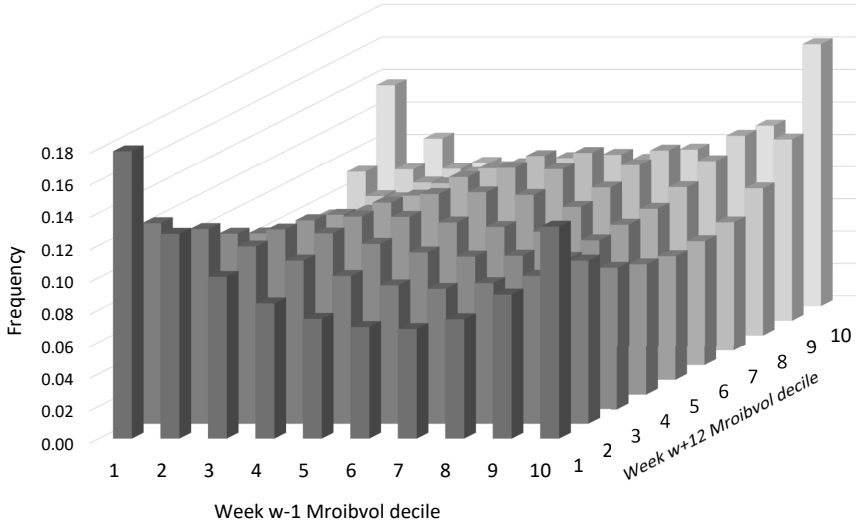
Tuttle, L. (2022). Accessing single dealer platforms (SDPs) in execution algorithms: Penny-wise and pound-foolish? *SEC White Paper*.

Vovchak, V. (2014). Liquidity and investment horizon. *Working Paper*.

Figures and Tables

Figure 3. Dynamics of *Mroibvol*: A Conditional Distribution. Panel A illustrates conditional distributions of $Mroibvol$ in week $w+12$ given $Mroibvol$ deciles in week $w-1$. Stocks are first sorted into deciles of $Mroibvol_{w-1}$. Within each deciles, stocks are then sorted into deciles of $Mroibvol_{w+12}$. The figure plots the relative frequencies of different $Mroibvol_{w+12}$ deciles at any given $Mroibvol_{w-1}$ decile. Panel B illustrates the analogous conditional distributions using simulated for a variable with $AR(1)$ structure $y_w = 0.8y_{w-1} + \epsilon_w$, with $\epsilon_w \sim N(0,1)$ and $y_0 = \epsilon_0$.

Panel A: Empirical joint distribution of $Mribvol$



Panel B: Simulated joint distribution of y

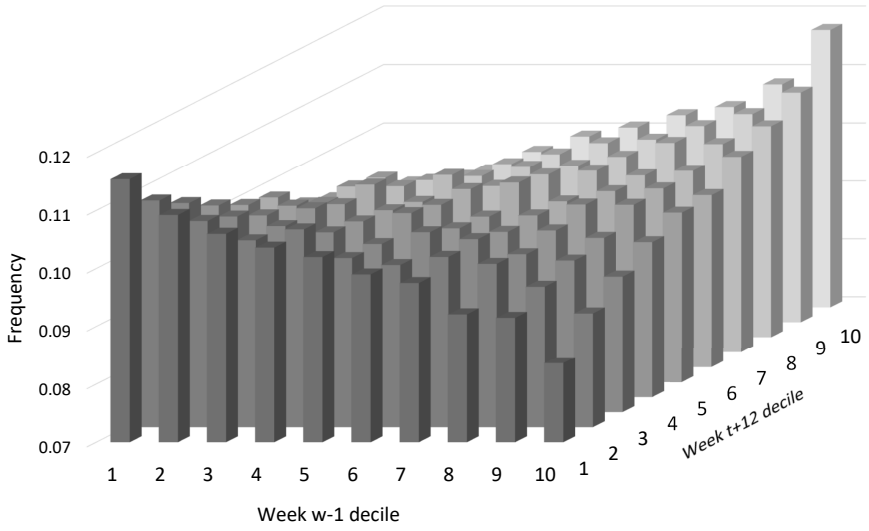
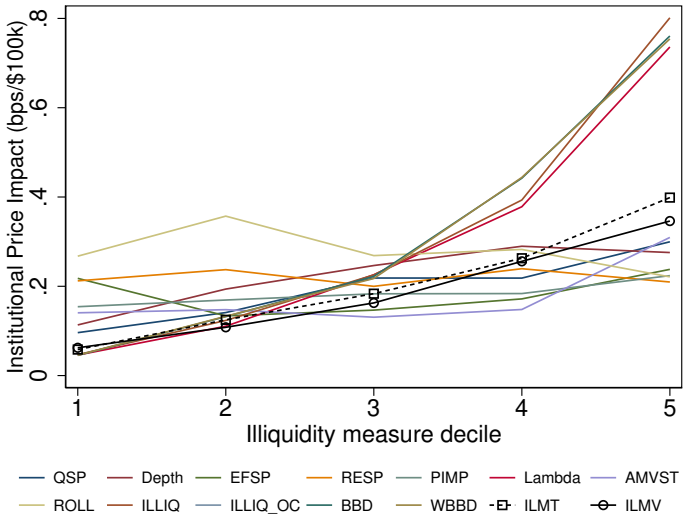


Figure 4. ***ILM*s, Standard Liquidity Measures, and Future Institutional Price Impacts.** The table reports on the cross-sectional relation between various liquidity measures constructed in month $m-2$ and realized, post-trade institutional price impacts, InPrIm, (in bps per $100k) constructed in month m. Liquidity measures include (1) quoted bid-ask spread (QSP); (2) quoted depth at best prices (Depth); (3) effective spreads (EFSP); (4) realized spreads (RESP); (5) price impacts (PIMP); (6) Kyle's lambda estimates (Lambda); (7) Amvist illiquidity measure (AMVST); (8) Roll measure of realized spreads (ROLL); (9 & 10) close-to-close and open-to-close Amihud measures (ILLIQ & ILLIQ_OC); (11 & 12) simple and volume-weighted trade-time liquidity measures (BBD & WBBD); (13 & 14) trade- and volume-based institutional liquidity measures (ILMT & ILMV). Each month, stocks are sorted into deciles of liquidity, with decile 1 (10) reflecting the most (least) liquid stocks, based on a given liquidity measure from month $m-2$. Month m InPrIm of the median stock in each liquidity decile is averaged across months by liquidity decile. This average is plotted against the respective liquidity decile. Panels A and B report results for liquidity deciles 1 through 5 and 6 through 10, respectively. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below $2.

Panel A: Illiquidity deciles 1–5



Panel B: Illiquidity deciles 6–10

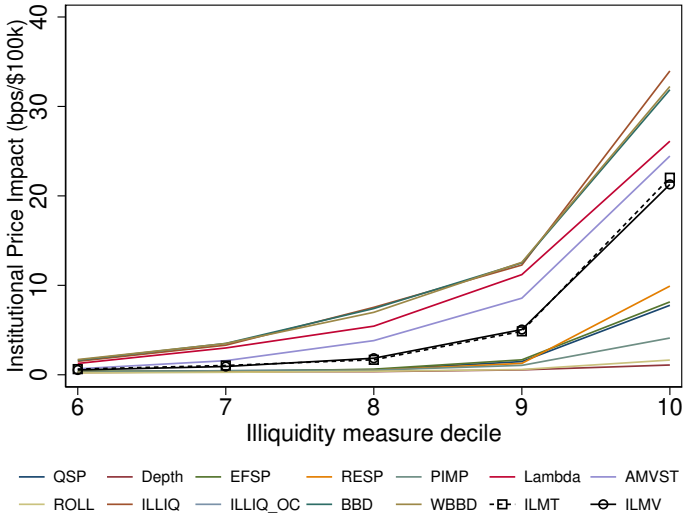


Figure 5. EUM Liquidity and Holding Horizon. This figure provides local polynomial estimates of equity under management (EUM) liquidity as a function of holding horizon. Holding weighted EUM liquidity, volatility, market capitalization, and institutional ownership are calculated for each manager. Every quarter, the residuals from regressing EUM liquidity on volatility, market capitalization, and institutional ownership are sorted into percentile statistics. Every quarter, manager-level holding horizons are calculated following Vovchak (2014) and sorted into percentile statistics. The figures present local polynomial estimates of residual EUM liquidity percentile statistics as functions of holding horizon percentile statistics. The sample includes all NMS common shares from January 2010 to December 2019. The sample for institutional price impacts (InPrIm) spans January 2010 through December 2019.

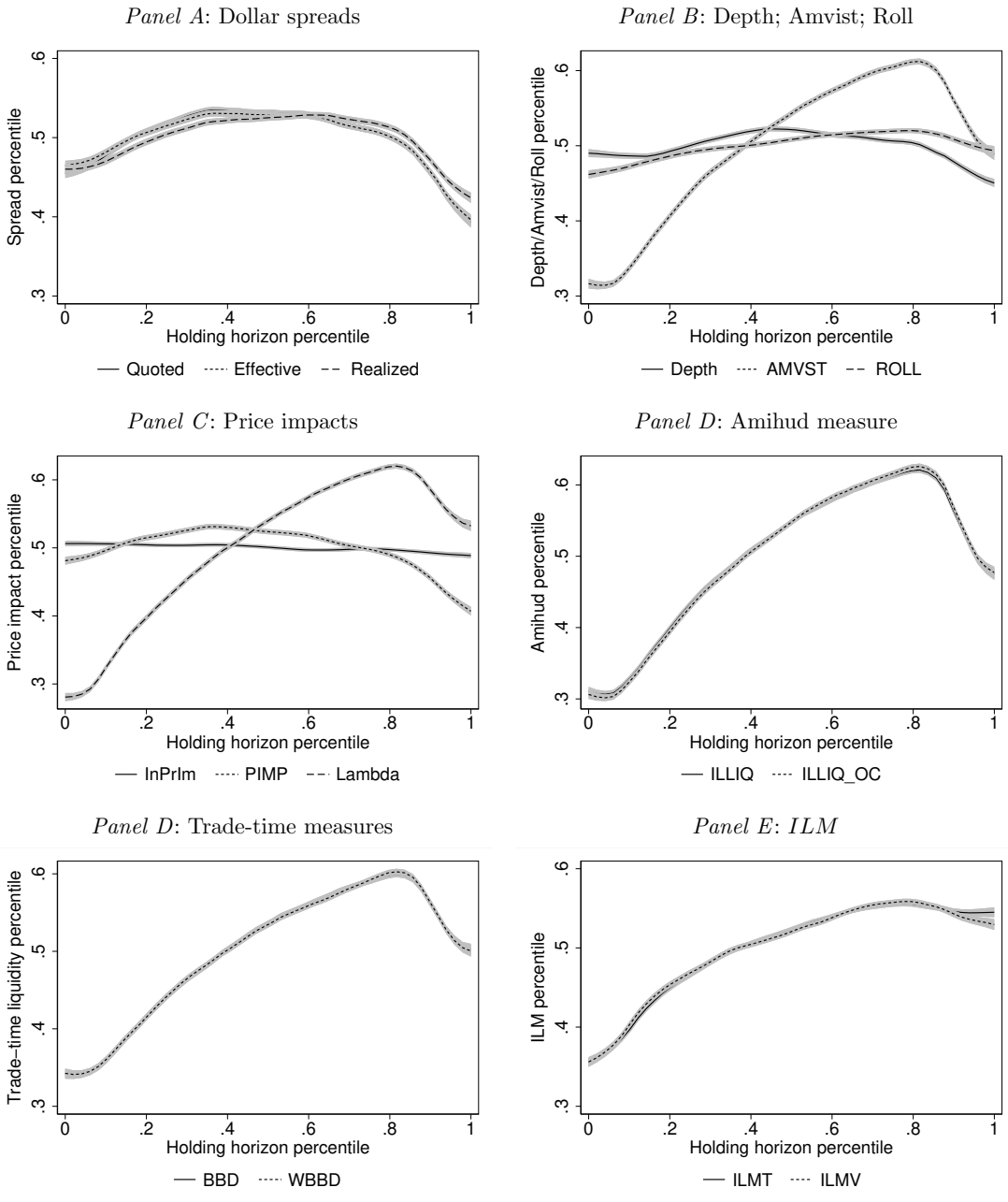


Table 1. Summary Statistics. Panel A reports (1) distributions of retail order types among all non-directed orders received by retail brokers; (2) distributions of retail order types, based on trade volume, among non-directed orders that are executed by wholesalers and receive PFOF; and (3) PFOF amount per 100 shares for different retail order types. All quantities are extracted from Charles Schwab, TD Ameritrade, and E*TRADE's 606 filing disclosures for the final quarter of 2020. When applicable, quantities reflect dollar-weighted averages across the top-5 wholesalers handling retail orders for the respective broker. Panel B reports summary statistics for daily measures of internalized order flows for our sample of NYSE-, AMEX-, and NASDAQ-listed common shares during the 2010–2014 period. $Mrbvol$ and $Mrsvol$ denote trading volumes for internalized trades classified as retail buy and retail sell, respectively. $Mrbtrd$ and $Mrstrd$ denote the number of internalized trades classified as retail buy and retail sell, respectively. $Mroibvol$ and $Mroibtrd$ then denote normalized imbalances in internalized retail order flow based on trading volume and trade frequency, respectively.

Panel A: Retail Orders Receiving Payment for Order Flow

	Charles Schwab			TD Ameritrade			E*TRADE		
	Non-directed orders (%)	Volume receiving PFOF (%)	PFOF (cents per 100 shares)	Non-directed orders (%)	Volume receiving PFOF (%)	PFOF (cents per 100 shares)	Non-directed orders (%)	Volume receiving PFOF (%)	PFOF (cents per 100 shares)
Market	52.9	57.2	9.0	18.8	44.7	12.0	49.3	53.7	19.9
Marketable limit	4.8	14.1	9.0	9.2	24.2	12.0	5.8	12.9	18.8
Non-marketable limit	33.8	21.1	29.6	31.9	21.2	33.5	35.0	18.0	29.3
Other order types	8.5	7.6	10.0	40.2	9.9	9.4	9.9	15.5	15.8
Total	100	100	–	100	100	–	100	100	–

Panel B: Internalized Retail Order Flow

	N	Mean	St. dev.	Median	Q1	Q3
$Mrbvol$	4,627,339	46,345	288,628	5,850	1,395	23,157
$Mrsvol$	4,627,339	46,249	270,718	6,333	1,559	24,346
$Mrbtrd$	4,627,339	108	389	23	6	79
$Mrstrd$	4,627,339	106	349	24	6	81
$Mroibvol$	4,627,339	−0.035	0.453	−0.025	−0.286	0.209
$Mrioibtrd$	4,627,339	−0.030	0.430	−0.008	−0.263	0.200
$Mroibvol > 0$	2,154,810	0.330	0.295	0.233	0.101	0.471
$Mroibvol < 0$	2,448,368	−0.357	0.301	−0.265	−0.522	−0.115
$Mroibtrd > 0$	2,088,865	0.321	0.282	0.232	0.111	0.435
$Mroibtrd < 0$	2,329,910	−0.347	0.290	−0.261	−0.500	−0.123

Table 2. Portfolios of *Mroibvol*: Contemporaneous Return, Liquidity, Institutional Trading, and Short Interest. The table presents the cross-sectional relationship between weekly $Mroibvol$ and the contemporaneous return, institutional trade, and liquidity outcomes. Outcome variables include (1) returns (close-to-close, intraday, and overnight returns, with a version of overnight returns shifted by one day); (2) liquidity (dollar and relative quoted spreads, depth, in shares, and abnormal off-exchange midpoint executions of larger trades); (3) institutional trading (actual trade imbalance, institutional price impact (in bps/$1m), and BJZZ-implied trade imbalance); and (4) short interest (% change in bi-weekly short interest). Each weekly cross-section is sorted into deciles of $Mroibvol$. The average of an outcome variable Y is calculated by $Mroibvol$ decile in each cross-section before the averages of mean-Y time-series are calculated. For short interest, bi-weekly relative % changes in short interest are constructed and $Mroibvol$ is aggregated over two-week periods, before forming $Mroibvol$ portfolios. Median short interest changes by $Mroibvol$ and stock size tercile, before averaging the time-series of medians.

	Deciles of internalized retail order flow imbalance ($Mroibvol$)									
	1	2	3	4	5	6	7	8	9	10
Mroibvol	−2.043	−1.132	−0.745	−0.467	−0.238	−0.033	0.173	0.417	0.763	1.607
Ratio of inside quote executions	0.158	0.135	0.126	0.123	0.121	0.122	0.120	0.122	0.132	0.162
Returns (%)										
Close-to-close return	−0.019	0.091	0.135	0.179	0.219	0.249	0.269	0.290	0.267	0.321
Intraday return	0.098	0.053	0.019	−0.005	−0.063	−0.118	−0.176	−0.210	−0.237	−0.138
Overnight return	−0.116	0.038	0.117	0.184	0.283	0.367	0.445	0.500	0.505	0.459
Next-day overnight return	−0.134	0.019	0.100	0.166	0.257	0.340	0.423	0.490	0.488	0.456
Institutional Trading										
Actual trade imbalance	0.277	0.265	0.264	0.247	0.238	0.228	0.212	0.212	0.202	0.172
Price impact	19.57	7.13	3.48	3.10	3.25	2.96	4.04	7.25	7.60	22.50
BJZZ-implied trade imbalance	−0.243	−0.257	−0.266	−0.270	−0.267	−0.250	−0.256	−0.252	−0.245	−0.221
Change in Short Interest (%)										
Small stocks	−2.58	−1.90	−1.38	−0.87	−0.61	0.22	0.16	0.70	1.21	2.25
Mid-sized stocks	−0.70	−0.54	−0.39	−0.10	−0.01	0.29	0.26	0.37	0.63	0.41
Large stocks	−1.16	−0.58	−0.72	−0.33	−0.25	−0.27	0.06	0.04	0.20	0.80
Liquidity										
Dollar quoted spread (¢)	8.9	6.8	5.8	5.4	5.3	5.7	5.4	5.5	6.4	9.3
Relative quoted spread (bps)	69	46	38	33	31	32	31	34	43	70
Ask-side depth	972	1,288	1,409	1,557	1,738	1,857	1,893	1,751	1,500	905
Bid-side depth	972	1,306	1,449	1,602	1,790	1,935	2,000	1,864	1,618	960
Large midpoint executions	0.79	0.89	0.94	0.98	1.00	1.04	1.07	1.06	1.03	0.99

Table 3. Internalized Retail Order Flow and the Cross-section of Next Week's Returns. This table presents estimates of the association between internalized retail order flow and the cross-section of the next week's returns (in percentage points). Daily returns of each stock are calculated based on the mid-points of best bid and ask prices at close as well as open prices, decomposing each day's close-to-close returns into intraday (open-to-close), and overnight (close-to-open) before aggregating each return type into weekly observations, denoted CCR_w, IDR_w,and ONR_w, respectively. According to equation (1), week w returns are regressed on week $w-1$'s internalized order flows ($Mroibvol_{w-1}$) and control variables including last week's return (CCR_{w-1}), last month's return (RET_{-1}), the return over the preceding five months ($RET_{(-7,-2)}$), volatility (VOLAT), and natural logs of turnover (ln(TO)), market capitalization (ln(Size)), and book-to-market ratio (ln(BM)). Estimates are based on Fama-Macbeth regressions, featuring Newey-West corrected standard errors with 6 lags. Sample includes NMS common shares from Jan 2010 – Dec 2014, excluding observations with previous month-end's closing price below $1. Numbers in brackets reflect t-statistics, and symbols ***, **, and * identify statistical significance at the 1%, 5%, and 10% type one errors, respectively.

Dependent Variable	CCR_w	ONR_w	IDR_w
Constant	0.0063	0.58***	−0.57**
	[0.02]	[4.58]	[−2.10]
$Mroibvol_{w-1}$	0.087***	0.12***	−0.029***
	[13.73]	[25.53]	[−4.41]
R_{w-1}	−0.021***	0.00090	−0.022***
	[−5.86]	[0.50]	[−7.07]
$RET_{(-1)}$	0.21	−0.19**	0.40**
	[1.14]	[−2.30]	[2.47]
$RET_{(-7,-2)}$	0.063	0.061**	0.0024
	[0.84]	[2.45]	[0.03]
ln(TO)	−0.037***	0.036***	−0.073***
	[−3.60]	[8.89]	[−8.16]
VOLAT	−6.44***	9.68***	−16.1***
	[−3.55]	[11.02]	[−10.03]
ln(Size)	0.020	−0.033***	0.053***
	[1.47]	[−5.31]	[4.39]
ln(BM)	0.058***	−0.038***	0.096***
	[2.73]	[−6.10]	[4.75]
Observations	3,330,408	3,330,408	3,330,408

Table 4. Portfolios of *Mroibvol* and Future Weekly Returns. The table presents the cross-sectional relationships between $Mroibvol$ and future weekly (%) returns. Each cross-section is sorted into portfolios (deciles) of $Mroibvol_{w-1}$ to calculate portfolio-specific averages of future close-to-close (CCR) returns in week $w+i$, with $i \in \{0, 1, 2, 3, 6, 9, 12, 24, 36, 39, 42, 45, 48, 51, 54, 57, 60\}$. The means of the time-series of portfolio future returns are presented by $Mroibvol$ decile.

	Deciles of $Mroibvol_{w-1}$									
Week	1	2	3	4	5	6	7	8	9	10
w	−0.02	0.09	0.13	0.18	0.22	0.25	0.27	0.29	0.27	0.32
$w+1$	0.13	0.15	0.14	0.15	0.15	0.14	0.17	0.16	0.21	0.34
$w+2$	0.14	0.16	0.17	0.16	0.16	0.15	0.16	0.17	0.21	0.31
$w+3$	0.17	0.20	0.18	0.18	0.17	0.17	0.17	0.18	0.23	0.29
$w+6$	0.19	0.17	0.19	0.18	0.16	0.16	0.18	0.18	0.21	0.26
$w+9$	0.14	0.16	0.16	0.13	0.13	0.12	0.10	0.11	0.15	0.19
$w+12$	0.15	0.12	0.11	0.10	0.08	0.07	0.07	0.09	0.12	0.18
$w+24$	0.21	0.18	0.19	0.15	0.14	0.13	0.13	0.15	0.16	0.22
$w+36$	0.22	0.21	0.20	0.17	0.15	0.13	0.14	0.15	0.17	0.20
$w+39$	0.16	0.17	0.16	0.14	0.13	0.11	0.10	0.10	0.13	0.14
$w+42$	0.18	0.15	0.13	0.13	0.12	0.11	0.09	0.08	0.12	0.15
$w+45$	0.19	0.17	0.15	0.14	0.12	0.10	0.09	0.10	0.12	0.14
$w+48$	0.14	0.13	0.11	0.09	0.07	0.05	0.06	0.04	0.06	0.10
$w+51$	0.13	0.10	0.12	0.07	0.02	0.02	0.01	0.03	0.04	0.07
$w+54$	0.08	0.10	0.08	0.08	0.04	0.01	0.00	-0.01	0.03	0.06
$w+57$	0.07	0.03	0.04	0.01	-0.01	-0.01	-0.01	0.02	0.03	0.05
$W+60$	0.08	0.07	0.04	0.01	0.00	0.00	-0.01	-0.02	0.00	0.00

Table 5. Institutional Liquidity Measures and Stock Characteristics. The table reports on the cross-sectional relation between ILMs and (1) three-factor Fama-French betas, (2) book-to-market ratios (BM), (3) natural log of market capitalizations (ln(Mcap)), (4) dividend yields (DYD), (5) idiosyncratic volatilities (IdVol), (6) previous month's returns ($RET_{(-1)}$), and (7) preceding returns from the prior 11 months ($RET_{(-12,-2)}$). Stock characteristics are computed from the prior month. Each weekly cross-section is sorted into ILM deciles. The average outcome variable is calculated by $ILMT$ decile in each cross-section before the average of the time-series is calculated. Panels A and B report the results for $ILMT$ and $ILMV$, respectively. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below $2.

Panel A: Trade-based Institutional Liquidity Measures ($ILMT$s) versus stock characteristics

	Weekly $ILMT$ deciles									
	1	2	3	4	5	6	7	8	9	10
Stock Characteristics:										
β^{mkt}	1.02	1.02	1.02	1.01	1.00	0.99	0.97	0.93	0.88	0.82
β^{hml}	0.73	0.73	0.73	0.73	0.74	0.75	0.76	0.77	0.78	0.79
β^{smb}	0.15	0.15	0.16	0.16	0.17	0.17	0.18	0.20	0.22	0.24
BM	0.64	0.64	0.65	0.65	0.66	0.67	0.68	0.72	0.76	0.80
ln(Mcap)	20.99	20.98	20.95	20.91	20.85	20.76	20.64	20.38	20.05	19.71
DYD	0.015	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.015	0.015
Id. Vol.	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.022	0.022
$RET_{(-1)}$	0.016	0.018	0.016	0.017	0.016	0.015	0.014	0.015	0.015	0.016
$RET_{(-12,-2)}$	0.19	0.19	0.19	0.19	0.19	0.18	0.17	0.16	0.15	0.14

Panel B: Volume-based Institutional Liquidity Measures ($ILMT$s) versus stock characteristics

	Weekly $ILMV$ deciles									
	1	2	3	4	5	6	7	8	9	10
Stock Characteristics:										
β^{mkt}	1.07	1.07	1.06	1.04	1.02	1.00	0.94	0.94	0.89	0.73
β^{hml}	0.71	0.71	0.72	0.73	0.73	0.75	0.74	0.79	0.82	0.77
β^{smb}	0.12	0.12	0.13	0.14	0.15	0.17	0.19	0.21	0.25	0.29
BM	0.62	0.62	0.63	0.63	0.64	0.65	0.70	0.70	0.74	0.87
ln(Mcap)	21.29	21.26	21.19	21.10	20.97	20.81	20.45	20.36	20.01	19.26
DYD	0.015	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.015	0.015
Id. Vol.	0.022	0.022	0.022	0.021	0.021	0.021	0.021	0.020	0.021	0.021
$RET_{(-1)}$	0.019	0.018	0.017	0.016	0.016	0.015	0.014	0.014	0.014	0.015
$RET_{(-12,-2)}$	0.21	0.21	0.20	0.19	0.19	0.18	0.16	0.16	0.15	0.13

Table 6. Persistence in the Institutional Liquidity Measures. The table reports on ILM's persistence. For $LIQ \in \{ILMT, ILMV\}$, monthly observations are regressed on monthly lagged observations from the preceding six months. Estimates are from Fama-MacBeth regressions that have Newey-West corrected standard errors with 6 lags. Both equally-weighted (EW) and value-weighted (VW) estimates, with weights being the previous month's market capitalization, are reported. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below $2. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

	ILMT		*ILMV*	
	EW	VW	EW	VW
Constant	0.0080***	0.0091***	0.0096***	0.0045***
	[5.81]	[6.14]	[7.84]	[5.80]
LIQ_{m-1}	0.40***	0.39***	0.43***	0.37***
	[69.77]	[33.97]	[83.17]	[49.29]
LIQ_{m-2}	0.19***	0.15***	0.19***	0.18***
	[54.73]	[14.43]	[55.50]	[31.86]
LIQ_{m-3}	0.13***	0.13***	0.13***	0.15***
	[37.56]	[14.46]	[47.16]	[31.93]
LIQ_{m-4}	0.078***	0.085***	0.068***	0.084***
	[19.72]	[10.00]	[21.83]	[10.64]
LIQ_{m-5}	0.070***	0.070***	0.060***	0.076***
	[22.27]	[9.70]	[23.77]	[15.89]
LIQ_{m-6}	0.090***	0.092***	0.087***	0.10***
	[39.04]	[14.33]	[31.25]	[16.66]
Observations	310,847	310,847	310,847	310,847

Table 7. Stock Liquidity and Institutional Holding Horizon. This table reports on the relation between the holding horizons of institutional investors and stock liquidity using different liquidity measures. Institutional investor turnover measures are constructed by stock and quarter as the weighted averages of turnover across the institutional investors holding a stock. For each stock, the weight assigned to an investor's turnover is the fraction held by the investor relative to the total amount held by institutional investors. Each quarter, investor-level holding horizon percentile statistics, "HH pctile", are defined as 1 minus institutional turnover percentile statistics across all the stocks held by an investor. In Panel A, for each stock j in quarter q, liquidity measure $LIQ_{j,q}$ is regressed on the holding horizon percentile statistic, return volatility, natural log of market capitalization, and institutional ownership from quarter $q-1$. Panel B reports on the relation between institutional turnover and liquidity, after orthogonalizing $ILMT$ and $ILMV$ with respect to existing liquidity measures and vice versa. Z_{ILMT} and Z_{ILMV}, respectively, are the residuals from regressing quarterly cross-sections of $ILMT$ and $ILMV$ on existing liquidity measures. Y_{ILMT} and Y_{ILMV}, respectively, are the residuals from regressing quarterly cross-sections of individual existing liquidity measures on $ILMT$ and $ILMV$. Z_{ILMT}, Z_{ILMV}, Y_{ILMT}, and Y_{ILMV} from quarter q are then regressed on institutional turnover, return volatility, natural log of market capitalization, and institutional ownership from quarter $q-1$. Institutional turnover coefficients are reported. Estimates are from Fama-MacBeth regressions that have Newey-West corrected standard errors with 6 lags. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below $2. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Stock liquidity and institutional turnover

	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD	ILMT	ILMV
HH pctile	**−7.07**	**0.12*** **	**−7.82*** **	**0.12*** **	**0.11** **	**0.0082**	**0.14*** **	**0.051*** **	**−0.00029**	**0.15*** **	**0.099*** **	**0.25**	**0.092** **	**0.093*** **	**0.12*** **
	[−0.81]	**[7.52]**	**[−6.50]**	**[3.26]**	**[2.66]**	**[0.40]**	**[4.24]**	**[6.92]**	**[−0.43]**	**[4.12]**	**[5.16]**	**[1.61]**	**[2.13]**	**[11.63]**	**[19.36]**
Volatility	435.6	−1.50***	239.9***	−0.26	−0.11	−0.23	5.61***	−0.25	0.19***	3.17***	2.15***	5.23***	2.79***	−2.73***	−3.60***
	[1.30]	[−7.40]	[3.94]	[−0.40]	[−0.15]	[−1.27]	[9.62]	[−1.49]	[17.36]	[3.75]	[4.54]	[4.85]	[6.17]	[−12.14]	[−19.65]
ln(Mcap)	0.88	−0.021***	3.94***	−0.015***	−0.0036	−0.011***	−0.15***	−0.020***	−0.0013***	−0.12***	−0.074***	−0.098***	−0.049***	−0.064***	−0.077***
	[1.13]	[−14.40]	[6.09]	[−10.84]	[−0.87]	[−2.79]	[11.19]	[−9.81]	[−17.40]	[13.03]	[13.62]	[−3.19]	[−5.11]	[−23.20]	[−46.22]
Ownership	−19.0	−0.089***	−18.0***	−0.095***	−0.13**	0.040	−0.56***	−0.12***	−0.0048***	−0.53***	−0.33***	−0.31***	−0.18***	−0.13***	−0.12***
	[−0.95]	[−7.55]	[−15.27]	[−4.37]	[−2.61]	[1.02]	[10.96]	[−8.17]	[−10.20]	[13.10]	[15.45]	[−9.81]	[−10.20]	[−27.37]	[−27.59]
R^2	**0.0061**	**0.092**	**0.026**	**0.095**	**0.021**	**0.011**	**0.36**	**0.027**	**0.13**	**0.11**	**0.14**	**0.18**	**0.18**	**0.61**	**0.63**
Obs.	28,679†	91,541	91,541	91,541	91,541	91,541	91,541	91,541	91,541	91,541	91,541	71,952††	71,952††	91,541	91,541

† The number of observations reflects the largest sample of ANcerno data available from 2010–2014.

†† The number of observations reflects the largest sample available for BBD and WBBD from 2010–2017.

Panel B: Stock liquidity and institutional turnover, ILM versus existing measures

Residual	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD
Z_{ILMT}	**0.10*** **	**0.053*** **	**0.092*** **	**0.055*** **	**0.087*** **	**0.090*** **	**0.078*** **	**0.089*** **	**0.092*** **	**0.086*** **	**0.082*** **	**0.090*** **	**0.090*** **
	[9.56]	**[10.18]**	**[12.25]**	**[9.47]**	**[8.28]**	**[10.02]**	**[13.96]**	**[11.08]**	**[13.39]**	**[11.52]**	**[12.15]**	**[12.56]**	**[12.96]**
R^2	0.60	0.54	0.61	0.54	0.60	0.61	0.41	0.60	0.57	0.55	0.52	0.53	0.53
Z_{ILMV}	**0.13*** **	**0.080*** **	**0.12*** **	**0.082*** **	**0.12*** **	**0.12*** **	**0.11*** **	**0.12*** **	**0.12*** **	**0.12*** **	**0.11*** **	**0.12*** **	**0.12*** **
	[17.39]	**[18.49]**	**[19.91]**	**[17.90]**	**[13.18]**	**[15.80]**	**[22.54]**	**[18.22]**	**[22.28]**	**[18.87]**	**[19.82]**	**[18.58]**	**[18.97]**
R^2	0.61	0.56	0.62	0.56	0.62	0.63	0.44	0.62	0.59	0.57	0.54	0.55	0.55
Y_{ILMT}	−5.60	**0.080*** **	**−7.17*** **	**0.085** **	**0.072***	0.014	−0.047***	−0.0081	−0.0018***	−0.069***	−0.029**	0.12	0.024
	[−0.59]	**[4.97]**	**[−4.82]**	**[2.44]**	**[1.86]**	[0.97]	[−3.13]	[−1.15]	[−3.25]	[−3.13]	[−2.23]	[0.95]	[0.66]
R^2	0.0026	0.025	0.022	0.025	0.0096	0.0069	0.13	0.029	0.086	0.024	0.031	0.057	0.058
Y_{ILMV}	−4.39	**0.070*** **	**−6.36*** **	**0.078** **	**0.069***	0.011	−0.082***	−0.013	−0.0018***	−0.099***	−0.049***	0.11	0.014
	[−0.47]	**[4.82]**	**[−4.36]**	**[2.24]**	**[1.77]**	[0.73]	[−4.52]	[−1.68]	[−3.46]	[−4.23]	[−3.51]	[0.85]	[0.41]
R^2	0.0026	0.025	0.020	0.025	0.0097	0.0065	0.14	0.022	0.092	0.030	0.038	0.065	0.065

Table 8. The Cross-Section of Expected Stock Returns and ILM. This table reports on the relation between alternative high-frequency liquidity measures and the cross-section of expected returns. In Panel A, equation (2) is estimated using liquidity measures ($LIQ_{j,m-2}$) constructed over 1-month horizons. Control variables include three-factor Fama-French betas ($\beta^{mkt}_{j,m-1}$, $\beta^{hml}_{j,m-1}$, $\beta^{smb}_{j,m-1}$), estimated using weekly observations from the two-year period ending in the final full week of month $m-1$, book-to-market ratio, ($\text{BM}_{j,m-1}$), natural log of market capitalization, ($\ln(\text{Mcap}_{j,m-1})$), dividend yield ($\text{DYD}_{j,m-1}$), defined as total dividends over the past 12 months divided by the share price at the end of month $m-1$, idiosyncratic volatility ($\text{IdVol}_{j,m-1}$), previous month's return ($RET_{(-1)}$), and preceding return from the prior 11 months ($RET_{(-12,-2)}$). Panel B replaces each high-frequency liquidity measure by the residuals of $ILMT$ and $ILMV$ with respect to each alternative liquidity measure, with residuals calculated separately for each monthly cross-section. Estimates are from Fama-MacBeth regressions that have Newey-West corrected standard errors with 6 lags. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$2. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Stock liquidity and the cross-section of expected returns

	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD	ILMT	ILMV
Constant	1.38	1.00	0.99	0.95	0.99	1.00	1.45*	0.99	1.41	1.13	1.00	1.68*	1.63*	−0.99	−1.54
	[1.08]	[1.11]	[1.14]	[1.06]	[1.15]	[1.15]	[1.73]	[1.16]	[1.60]	[1.30]	[1.13]	[1.93]	[1.87]	[−0.77]	[−1.13]
Liquidity	**0.029***	**0.0057**	**−0.00**	**0.13**	**0.049**	**−0.034**	**−0.11**	**0.043**	**−8.24*****	**−0.015**	**0.050**	**−0.070**	**−0.055**	**1.20*****	**1.27*****
	[1.91]	**[0.05]**	**[−0.84]**	**[0.78]**	**[0.63]**	**[−0.33]**	**[−1.53]**	**[0.35]**	**[−3.47]**	**[−0.45]**	**[0.56]**	**[−0.56]**	**[−0.28]**	**[2.91]**	**[3.11]**
β^{mkt}	−0.023	−0.15	−0.15	−0.15	−0.15	−0.15	−0.16	−0.16	−0.15	−0.16	−0.15	−0.17	−0.17	−0.070	−0.043
	[−0.06]	[−0.75]	[−0.75]	[−0.74]	[−0.74]	[−0.75]	[−0.78]	[−0.75]	[−0.71]	[−0.76]	[−0.75]	[−0.71]	[−0.70]	[−0.36]	[−0.23]
β^{hml}	−0.15	−0.098	−0.097	−0.097	−0.098	−0.098	−0.096	−0.097	−0.10	−0.098	−0.096	−0.064	−0.064	−0.11	−0.12
	[−1.02]	[−0.83]	[−0.82]	[−0.82]	[−0.82]	[−0.82]	[−0.81]	[−0.82]	[−0.88]	[−0.82]	[−0.81]	[−0.47]	[−0.47]	[−0.92]	[−0.98]
β^{smb}	0.12	0.063	0.062	0.064	0.062	0.061	0.053	0.064	0.060	0.052	0.060	0.057	0.061	0.10	0.11
	[1.28]	[0.84]	[0.82]	[0.86]	[0.83]	[0.81]	[0.69]	[0.85]	[0.79]	[0.68]	[0.80]	[0.67]	[0.71]	[1.44]	[1.58]
BM	0.22	0.0056	0.0059	0.0058	0.0056	0.0052	−0.0015	0.0044	0.0088	0.0073	0.0023	0.055	0.054	0.0030	0.0043
	[1.52]	[0.11]	[0.12]	[0.12]	[0.11]	[0.11]	[−0.03]	[0.09]	[0.18]	[0.15]	[0.05]	[0.71]	[0.69]	[0.06]	[0.09]
ln(Mcap)	0.0048	0.022	0.023	0.023	0.023	0.022	0.0024	0.022	0.0055	0.016	0.022	−0.0054	−0.0030	0.097*	0.12**
	[0.09]	[0.59]	[0.62]	[0.62]	[0.63]	[0.61]	[0.07]	[0.62]	[0.15]	[0.44]	[0.59]	[−0.15]	[−0.08]	[1.89]	[2.15]
DYD	0.35	−0.049	−0.062	−0.050	−0.066	−0.075	−0.070	−0.053	−0.077	−0.088	−0.086	0.11	0.11	−0.13	−0.11
	[0.31]	[−0.09]	[−0.11]	[−0.09]	[−0.12]	[−0.13]	[−0.12]	[−0.09]	[−0.14]	[−0.15]	[−0.15]	[0.17]	[0.17]	[−0.23]	[−0.20]
Id. Vol.	−0.16**	−0.23***	−0.23***	−0.23***	−0.23***	−0.23***	−0.23***	−0.23***	−0.22***	−0.23***	−0.24***	−0.23***	−0.23***	−0.22***	−0.21***
	[−2.47]	[−4.75]	[−4.78]	[−4.75]	[−4.76]	[−4.75]	[−4.62]	[−4.77]	[−4.51]	[−4.69]	[−4.65]	[−4.01]	[−4.05]	[−4.54]	[−4.46]
RET_{-1}	−0.74	−0.38	−0.39	−0.38	−0.37	−0.36	−0.36	−0.37	−0.39	−0.33	−0.35	−0.42	−0.43	−0.44	−0.48
	[−1.04]	[−0.81]	[−0.82]	[−0.81]	[−0.78]	[−0.77]	[−0.75]	[−0.79]	[−0.82]	[−0.70]	[−0.74]	[−0.79]	[−0.80]	[−0.93]	[−1.02]
$RET_{(-12,-2)}$	0.35*	0.21	0.21	0.21	0.21	0.21	0.18	0.21	0.21	0.20	0.20	0.21	0.21	0.27*	0.28*
	[1.80]	[1.39]	[1.39]	[1.39]	[1.39]	[1.40]	[1.14]	[1.38]	[1.37]	[1.32]	[1.30]	[1.11]	[1.13]	[1.76]	[1.81]
Observations	128,135†	340,227	340,227	340,227	340,227	340,227	339,681	340,225	340,227	340,225††	340,225††	277,750†††	277750†††	340,227	340,227

Panel B: Loadings of ILMs in the cross-section of expected returns after orthogonlization relative to other liquidity measures

	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD	ILMT	ILMV
ILMT residual	0.10	1.22***	1.19***	1.15***	1.18***	1.20***	1.30***	1.20***	1.38***	1.27***	1.13**	1.14**	1.12**	-	-
	[0.19]	[3.51]	[2.92]	[3.27]	[2.85]	[2.90]	[2.85]	[2.77]	[3.35]	[2.90]	[2.48]	[2.18]	[2.17]		
ILMV residual	0.055	1.31***	1.24***	1.25***	1.25***	1.28***	1.34***	1.25***	1.40***	1.31***	1.19***	1.17**	1.15**	-	-
	[0.11]	[3.85]	[3.11]	[3.60]	[3.05]	[3.14]	[3.05]	[2.98]	[3.45]	[3.11]	[2.76]	[2.30]	[2.29]		

† The number of observations reflects the largest sample of ANcerno data available from 2010–2014.

†† The number of observations reflects the largest sample available for ILLIQ and ILLIQ_OC.

††† The number of observations reflects the largest sample available for BBD and WBBD from 2010–2017.

Table 9. The Cross-Section of Expected Stock Returns and ILM: Robustness Tests. This table reports on the robustness of the relation between between our institutional liquidity measures and the cross-section of expected stock returns. Equation (2) is estimated using institutional liquidity measures ($LIQ_{j,m-2}$) constructed over 1-month horizons. Control variables include three-factor Fama-French betas ($\beta^{mkt}_{j,m-1}$, $\beta^{hml}_{j,m-1}$, $\beta^{smb}_{j,m-1}$), estimated using weekly observations from the two-year period ending in the final full week of month $m-1$, book-to-market ratio ($\text{BM}_{j,m-1}$), natural log of market capitalization ($\ln(\text{Mcap}_{j,m-1})$), dividend yield ($\text{DYD}_{j,m-1}$), defined as total dividends over the past 12 months divided by the share price at the end of month $m-1$, idiosyncratic volatility ($\text{IdVol}_{j,m-1}$), previous month's return ($RET_{(-1)}$), and preceding return from the prior 11 months ($RET_{(-12,-2)}$). Panel A reports on the robustness of the results to (1) estimating coefficients using panel regressions with date and stock fixed effects and date-stock double-clustered standard errors, (2) weighting observations (by size or according to Asparouhova et al. 2010) to correct for microstructure noise, (3) excluding firms with the smallest 20% market capitalization, (4) excluding stocks in the bottom 10% of the ratio of sub-penny volume in total volume; and (5) excluding stocks in the top or bottom 10% of the respective ILM. Stocks whose previous month-end's closing price is below $p_{min} \in \{\$1, \$2, \$5\}$ are excluded. Panel B reports on the robustness of the estimates in equation (2) to listing exchange. Observations are weighted according to Asparouhova et al. (2010) after excluding stocks whose previous month-end's closing price is below \$1 and stocks falling in the bottom 10% of the ratio of sub-penny volume in total volume. Estimates are from Fama-MacBeth regressions that have Newey-West corrected standard errors with 6 lags. The sample includes NMS common shares from January 2010 to December 2019. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Robustness to estimation method and sample selection

Robustness specification	$ILMT$			$ILMV$		
	Price > \$1	Price > \$2	Price > \$5	Price > \$1	Price > \$2	Price > \$5
Panel regressions + stock & date FEs	1.20**	1.17**	0.55	1.54***	1.27***	0.80*
+ double-clustered S.E.	[2.18]	[2.25]	[1.16]	[2.98]	[2.64]	[1.85]
Asparouhova et al. (2010)	1.19**	1.18***	0.66*	1.35***	1.24***	0.88**
	[2.45]	[2.72]	[1.88]	[2.80]	[2.83]	[2.43]
Asparouhova et al. (2010)	0.99**	0.95**	0.62*	1.10**	1.06**	0.84**
+ top 80% market capitalization	[2.38]	[2.41]	[1.74]	[2.52]	[2.57]	[2.30]
Asparouhova et al. (2010)	1.33***	1.34***	0.86**	1.51***	1.41***	1.09***
+ low sub-penny volume stocks excluded	[2.64]	[2.98]	[2.37]	[3.02]	[3.09]	[2.89]
Size-weighted estimation	1.50**	1.52**	1.53**	0.38	0.38	0.36
	[2.38]	[2.39]	[2.35]	[0.73]	[0.72]	[0.67]
Stocks in top and bottom	2.42***	2.35***	1.33***	1.77***	1.62***	1.35***
10% of ILM excluded	[2.92]	[3.29]	[2.72]	[2.96]	[2.93]	[2.92]

Panel B: Robustness to estimation by listing exchange

	$ILMT$		$ILMV$	
	NYSE/AMEX	NASDAQ	NYSE/AMEX	NASDAQ
Asparouhova et al. (2010)	0.83	1.11**	1.17**	1.25**
+ Price > \$1	[1.57]	[2.14]	[2.15]	[2.55]
Asparouhova et al. (2010) + Price > \$1	1.04*	1.20**	1.43**	1.36***
+ low sub-penny volume stocks excluded	[1.90]	[2.29]	[2.48]	[2.73]

Table 10. Liquidity Alphas. This table presents three-factor alphas conditional on our liquidity measures. Panels A, B, and C report results based on NMS-listed common shares using CRSP breakpoints and equally-weighted portfolio returns. Panels D, E, and F report results based on the NMS-listed common shares, after removing stocks with the smallest 20% market capitalization at the end-of-last-month, using NYSE breakpoints and value-weighted portfolio returns. Stocks in each monthly cross-section are sorted into ten ILM portfolios (deciles). Monthly portfolio returns are averages of monthly stock returns in the portfolio. The time-series feature 118 months. The time-series returns of each portfolio (after subtracting the 1-month Treasury-bill rate) including the long-short portfolio are then regressed on Fama-French three factors. The resulting intercepts represent three-factor alphas. The sample period is from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below $p_{min} \in \{\$1, \$2, \$5\}$. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: CRSP breakpoints, \$1 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.32***	−0.34***	−0.19**	−0.17	−0.23***	−0.24*	−0.032	0.089	0.38**	0.64***	0.96***
	[−2.77]	[−3.82]	[−2.13]	[−1.58]	[−2.80]	[−1.83]	[−0.30]	[0.63]	[2.48]	[4.25]	[4.30]
$ILMV$	−0.63***	−0.44***	−0.25***	−0.25***	−0.11	0.00096	−0.027	0.32***	0.32**	0.64***	1.27***
	[−4.28]	[−4.40]	[−2.88]	[−3.56]	[−1.07]	[0.01]	[−0.28]	[2.85]	[2.10]	[4.76]	[5.49]

Panel B: CRSP breakpoints, \$2 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.30***	−0.33***	−0.21**	−0.062	−0.18**	−0.14	0.023	0.11	0.34**	0.62***	0.93***
	[−2.70]	[−4.05]	[−2.17]	[−0.82]	[−2.26]	[−1.33]	[0.27]	[0.92]	[2.54]	[4.48]	[4.33]
$ILMV$	−0.58***	−0.33***	−0.23***	−0.25***	−0.084	0.091	0.041	0.28***	0.31**	0.63***	1.20***
	[−3.97]	[−3.86]	[−2.76]	[−3.68]	[−0.92]	[1.12]	[0.59]	[3.37]	[2.26]	[4.97]	[5.09]

Panel C: CRSP breakpoints, \$5 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.29***	−0.24***	−0.14*	0.053	0.019	−0.0071	0.12	0.28***	0.38***	0.65***	0.95***
	[−2.66]	[−2.89]	[−1.98]	[0.78]	[0.26]	[−0.11]	[1.26]	[2.84]	[3.49]	[4.72]	[4.30]
$ILMV$	−0.43***	−0.21***	−0.14**	−0.11	0.0080	0.048	0.19***	0.37***	0.43***	0.68***	1.10***
	[−3.35]	[−2.64]	[−2.16]	[−1.54]	[0.10]	[1.01]	[2.86]	[4.65]	[4.02]	[5.32]	[4.82]

Continued on next page

Table 10 − *continued from previous page*

Panel D: NYSE breakpoints, largest 80% market capitalization, $1 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.10	−0.0096	−0.0039	0.0073	0.10	0.23**	0.19**	0.26*	0.15*	0.47***	0.58***
	[−1.58]	[−0.10]	[−0.05]	[0.06]	[0.90]	[2.61]	[2.37]	[1.87]	[1.76]	[7.07]	[6.09]
$ILMV$	−0.084	0.085	−0.026	−0.026	0.12	0.069	0.19*	0.25***	0.32**	0.32***	0.41***
	[−1.41]	[1.20]	[−0.29]	[−0.29]	[1.17]	[0.65]	[1.87]	[3.40]	[2.42]	[3.12]	[4.05]

Panel E: NYSE breakpoints, largest 80% market capitalization, $2 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.099	−0.017	−0.015	−0.0083	0.14	0.17	0.22**	0.24*	0.17*	0.48***	0.58***
	[−1.51]	[−0.18]	[−0.20]	[−0.06]	[1.29]	[1.64]	[2.51]	[1.77]	[1.93]	[7.12]	[6.15]
$ILMV$	−0.086	0.086	−0.016	−0.030	0.11	0.071	0.17	0.26***	0.28**	0.37***	0.46***
	[−1.43]	[1.18]	[−0.19]	[−0.32]	[1.12]	[0.67]	[1.64]	[3.33]	[2.24]	[3.63]	[4.69]

Panel F: NYSE breakpoints, largest 80% market capitalization, $5 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.10	−0.041	0.024	0.0047	0.20**	0.082	0.33***	0.17	0.10	0.53***	0.63***
	[−1.58]	[−0.46]	[0.29]	[0.03]	[2.01]	[0.77]	[3.46]	[1.34]	[1.04]	[7.20]	[6.17]
$ILMV$	−0.091	0.11	−0.060	−0.0087	0.11	0.086	0.22**	0.21**	0.28***	0.34***	0.43***
	[−1.52]	[1.38]	[−0.68]	[−0.10]	[1.22]	[0.81]	[2.47]	[2.25]	[2.65]	[2.91]	[4.27]

Internet Appendix

A Economics of Retail Order Internalization

A.1 Wholesaler Incentives, *Mroib*, and Institutional Liquidity

In this section, we provide a setting to illustrate the economic incentives underlying a wholesaler's decisions about which retail orders to internalize, and the consequences for $Mroib$. We focus on a setting where the wholesaler faces variable costs of internalization due to the possibility of internalizing both marketable and non-marketable orders. Similar economic considerations arise in a framework where internalization of marketable orders is sometimes more costly as a result of inside quote hidden liquidity (due to the Manning rule).

Suppose that the public information value of a share is V, and there is a four tick spread. Thus, the bid is $\$(V-2t)$ and the ask is $\$(V+2t)$. The distribution of retail orders routed by the broker-dealer to a wholesaler is given by

- n^s_{-2} marketable sell orders at $\$(V-2t)$
- n^s_{-1} limit sell orders at $\$(V-t)$
- n^s_0 limit sell orders and n^b_0 limit buy orders at $\$V$
- n^b_1 limit buy orders at $\$(V+t)$
- n^b_2 marketable buy orders at $\$(V+2t)$

To illustrate the economics, suppose there is more retail sell interest than retail buy interest so that $n^s_{-j} \geq n^b_j$, for $j = 0, 1, 2$, and we define $\Delta_j = n^s_{-j} - n^b_j \geq 0$. To reduce the number of cases that we need to enumerate, we assume that (a) $n^s_{-2} \leq n^b_2 + n^b_1$, and (b) $n^s_{-2} + n^s_{-1} \leq n^b_2 + n^b_1 + n^b_0$. Qualitatively similar implications obtain when these assumptions do not hold.

The wholesaler chooses whether to internalize a retail order in return for giving the broker-dealer PFOF, or to reroute it directly to an exchange, in which case all rebates (or fees) go to the retail broker, where the rebate for liquidity-making limit orders exceeds that for liquidity-taking market orders.[60] The broker-dealer obtains $PFOF_j$ in return for outsourcing the execution of a

[60] A third possibility in practice is that the wholesaler can post similarly-priced orders out of its own inventory on an exchange, and fill the order received if its proprietary order is executed on an exchange, where upon execution, the wholesaler internalizes the retail order and pays PFOF.

type j order to the wholesaler.

Price improvement of $PI_M > 0$ is offered to marketable orders in order to satisfy best execution duties. For simplicity, we assume that fraction $\alpha_{NM} \geq 0$ of non-marketable orders receive price improvement of $PI_{NM} > 0$. As we show, a large share of trade executions with sub-penny price improvements are inside the NBBO, indicating that α_{NM} is non-trivial. To ease presentation, we assume that the total PFOF plus PI offered is less than half a tick, so that it is profitable to intermediate buy and sell orders than are one tick apart.

It is costly for the wholesaler to hold inventory that deviates by q from its preferred inventory level of 0. The notion that a market-maker has "preferred" inventory positions dates back to Amihud and Mendelson (1980).[61] We assume that these costs rise convexly in q, i.e., $c(q) - c(q-1)$ is strictly increasing in q, consistent with risk-averse liquidity providers as in Grossman and Miller (1988) or Campbell, Grossman, and Wang (1993), where $c(1) - c(0)$ is assumed to be less than the expected liquidity rebate, consistent with tiny deviations from optimal inventory levels not being that costly.

We first highlight the economic forces for balanced levels of $Mroib$ in the absence of institutional liquidity demand. When a wholesaler is not "pinged" by an institution, it is strictly profitable for the wholesaler to internalize marketable sell orders and limit sell orders at $\$(V-t)$ simultaneously with marketable buy orders and limit buy orders at $\$(V+t)$, as the PFOF plus PI paid is less than the profit obtained by intermediating these orders. Thus, at least $min\{n^s_{-2}+n^s_{-1}, n^b_2+n^b_1\} = n^b_2+n^b_1$ is filled on each side by the wholesaler's internalization. The BJZZ algorithm identifies the subset of those internalized orders that receives price improvement, which comprise a total of $2(n^b_2+\alpha_{NM}n^b_1)$.

After filling these orders, the distribution of the remaining retail orders is given by

- 0 marketable sell orders at $\$(V-2t)$
- $n^s_{-2} + n^s_{-1} - (n^b_2 + n^b_1)$ limit sell orders at $\$(V-t)$
- n^s_0 limit sell orders and n^b_0 limit buy orders at $\$V$
- 0 limit buy orders at $\$(V+t)$
- 0 marketable buy orders at $\$(V+2t)$

[61] Other early studies suggesting or modeling the existence of such inventory positions include Smidt (1971), Barnea and Logue (1975), Stoll (1976), Ho and Stoll (1982), and Grossman and Miller (1988), among others.

Next observe that it is optimal for the wholesaler to internalize some of the remaining limit sell orders at $\$(V-t)$ by holding inventory, stopping at the inventory imbalance of q^* where

$$\begin{aligned} t-(c(q^*)-c(q^*-1)) &\geq t-PFOF_1-PFOF_0-2\alpha_{NM}PI_1 \\ &> t-(c(q^*+1)-c(q^*)). \end{aligned}$$

That is, the wholesaler stops internalizing orders when the marginal profit from internalizing by holding more unbalanced inventory would be less than that from simultaneously filling a non-marketable limit sell order at $\$(V-t)$ and a non-marketable limit buy order at $\$V$. Again, BJZZ's algorithm identifies fraction α_{NM} of these orders.

When $n^s_{-2}+n^s_{-1}-(n^b_2+n^b_1)>q^*$, the wholesaler fills the remaining limit sell orders at $\$(V-t)$ with limit buy orders at $\$V$. The dealer then submits all remaining limit orders[62] at $\$V$ to exchanges. Thus, absent institutional liquidity demand, for $n^s_{-2}+n^s_{-1}\leq n^b_2+n^b_1+q^*$, internalization order imbalances identified by the BJZZ algorithm equal

$$|Mroibvol| = \frac{(n^s_2+\alpha_{NM}n^s_1)-(n^b_{-2}+\alpha_{NM}n^b_{-1})}{n^b_2+\alpha_{NM}n^b_1+n^s_{-2}+\alpha_{NM}n^s_{-1}} = \frac{\Delta_2+\alpha_{NM}\Delta_1}{n^b_2+n^s_{-2}+\alpha_{NM}(n^b_1+n^s_{-1})}.$$

$|Mroibvol|$ reaches a maximum at $n^s_{-2}+n^s_{-1}=n^b_2+n^b_1+q^*$, where substituting for $\Delta_1=q^*-\Delta_2$ yields

$$|Mroibvol| = \frac{\alpha_{NM}q^*+(1-\alpha_{NM})\Delta_2}{2(n^b_2+\alpha_{NM}n^b_1)+\alpha_{NM}q^*+(1-\alpha_{NM})\Delta_2}.$$

For $n^s_{-2}+n^s_{-1}>n^b_2+n^b_1+q^*$, $|Mroibvol|$ falls with further increases in n^s_{-1}, as sell orders at $\$V-t$ are crossed with buy orders at $\$V$, while the denominator rises due to the "crossing" of the fraction α_{NM} receiving price improvement. Thus, if $\alpha_{NM}=1$, then a peak of

$$|Mroibvol| = \frac{q^*}{2(n^b_2+n^b_1)+q^*}$$

is reached, and if $\alpha_{NM}=0$, then the peak is

$$|Mroibvol| = \frac{q^*-\Delta_1}{2n^b_2+q^*-\Delta_1}$$

Thus, with no institutional liquidity demand, we predict that internalization of retail orders should

[62] That is, the n^s_0 limit sell orders, and the $n^b_0-q^*-(n^s_{-2}+n^s_{-1}-(n^b_2+n^b_1))$ remaining limit buy orders.

be roughly balanced.

Now suppose there is significant institutional liquidity demand. Such demand, when non-zero, is likely large relative to retail order flow, reflecting the much larger positions that institutions take, and the fact that there is little point for an institution to ping a wholesaler for a small position. To highlight how institutional demand changes $Mroib$ measures, suppose now that there is extensive institutional sell demand in the setting above, where previously there were relatively small negative (sell) retail trade imbalances.

Internalized order flow is an expensive source of liquidity for institutions. To see why, first note the straightforward direct effect—an institution seeking to sell shares must compensate a wholesaler for the profits that the wholesaler would otherwise obtain by internalizing retail sell orders. More subtly, an institution must also compensate a wholesaler for the foregone possibility of using the internalized retail buy orders to profitably fill retail sell orders without distorting the wholesaler's inventory—retail buy orders that are used to fill institutional sell orders cannot be used to fill retail sell orders. Finally, a wholesaler may have some bargaining power in negotiations with institutions. This logic implies that an institution interested in selling shares on an SDP must compensate the wholesaler via a combination of a low purchase price p_s and SDP access fees.

To begin suppose that the institution seeks to sell more than $n_2^b + n_1^b + n_0^b + q_s^*$ where

$$\begin{aligned} V - p_s - (c(q_s^*) - c(q_s^* - 1)) &\geq 0 \\ &> V - p_s - (c(q_s^* + 1) - c(q_s^*)). \end{aligned}$$

Then a wholesaler will internalize the retail buy orders received ($n_2^b + n_1^b + n_0^b$) to fill the institution's sell orders, and continue to fill them via increasing its inventory only up to the point ($n_2^b + n_1^b + n_0^b + q_s^*$) where the marginal profit from internalization exceeds the marginal increase in inventory costs. Now, all retail sell orders are rerouted to other trading venues so that, rather than being negative, $Mroibvol$ takes on its maximum value of one.

From this point, as one reduces institutional sell demand, one eventually reaches the level ($n_2^b + n_1^b + n_0^b + q_s^*$) below which a wholesaler now fills all of the institution's orders. To do this, a wholesaler uses all retail buy orders while distorting its inventory to the minimum extent needed, and still reroutes all retail sell orders to trading venues. Thus, on this range, the marginal order

is accommodated out of inventory, so $Mroibvol = 1$, remaining maximally tilted in the opposite direction of true retail order flow imbalance, $\frac{\sum_j \Delta_j}{\sum_j (n_j^b + n_{-j}^s)} < 0$.

With further reductions, one reaches a level of institutional sell demand at which the marginal inventory cost just falls below the profit from filling a marketable retail sell order. At this point, a wholesaler starts to internalize marketable retail sell orders, causing $|Mroibvol|$ to begin to fall, as first more attractive retail sell limit orders are internalized, and then limit buy orders at $\$V$ are rerouted to other trading venues instead of being internalized.

Taken together the observations with and without institutional liquidity demand reveal that (i) small *Mroib* imbalances are an indication of the absence or near absence of net institutional demand, while (ii) very large *Mroib* imbalances indicate unbalanced net institutional liquidity demand with the opposite sign of *Mroib*.

A.2 Minimum Tick Sizes and Internalization

In this section, we exploit the design of the Tick Size Pilot to establish that variation in $Mroibtrd$ and $Mroibvol$ reflects the internalization decisions of wholesalers. We first examine the response in a wholesaler's appetite to internalize, proxied by the extent of off-exchange sub-penny BJZZ-identified trading volume, to a shock in the profitability of wholesaler liquidity provision. More importantly, we also analyze the effect of a shock to the cost of internalization on imbalances in $Mroibtrd$ and $Mroibvol$. This analysis allows us to link wholesaler cost-benefit considerations to their choices of which retail orders to internalize.

The SEC implemented the Tick Size Pilot program (TSP) on October 3, 2016. This program offered an experimental design for studying the causal impact of the minimum tick size on trading outcomes. The program included 2,400 securities. To ensure that stocks were randomly assigned to control and treatment groups, stocks were sorted into 27 categories based on share price, market-capitalization, and trading volume terciles. Across these categories, stocks were randomly assigned to three treatment groups of 400 stocks each. Treated stocks in Test Group 1 were subject to a minimum quoting requirement of 5¢ but could trade at price increments of 1¢—the *quote rule* (Rindi and Werner (2019)). Treated stocks in Test Groups 2 and 3 were subject to a minimum quoting requirement of 5¢ and had to trade at price increments of 5¢—the *trade rule* (Rindi and Werner (2019)). Test Group 3 stocks were also subject to a Trade-At Prohibition provision that

effectively prevented sub-penny off-exchange execution prices, rendering test Group 3 irrelevant for our study (see Hu and Murphy (2022)).[63]

A key exception to the minimum tick size applied to retail trades. Although retail trades are quoted using the minimum tick size, they could be executed at sub-penny prices off-exchange. While TSP did not restrict the magnitudes of PI for test Group 1, the program imposed a minimum PI of 0.5¢ for off-exchange retail order executions of Test Group 2 stocks, raising the cost of internalizing orders in test Group 2 stocks above that for control and test Group 1 stocks.[64] This key difference provides an opportunity to examine the causal impacts of internalization costs on $Mroib$ imbalances.

BJZZ's algorithm is designed to detect sub-penny execution prices in a 1¢ tick size regime, but it can be scaled to detect sub-tick execution prices in any tick size regime. To do this for Test Group 2, after activation of the Trade Rule, we re-scale the algorithm's command that classifies trades according to small vs. large sub-penny increments by a factor of 5: in BJZZ's notation, we replace "$Z_{jt} = 100 * \text{mod}(P_{jt}, 0.01)$" by "$Z^5_{jt} = 20 * \text{mod}(P_{jt}, 0.05)$", where Z^5_{jt} is the *sub-tick* execution price (P_{jt}) increment for a 5¢ tick size. With this scaling, $Z^5_{it} \in [0, 1]$ and transactions can be classified into retail buy and retail sell trades as in Section 4.

The TPS provides an ideal setting to study the economics of retail flow internalization by wholesalers since the experiment raises (i) the profitability of off-exchange liquidity provision in all test groups (Rindi and Werner 2018); and (ii) the costs of internalization in test Group 2. These impacts let us conclude that variation in $Mroibtrd$ and $Mroibvol$ is determined by wholesaler decisions to internalize specific retail orders. We use the following Difference-in-Difference (DiD) methodology to examine the causal impact of a tick size change:

$$X_{j,d} = b^g_0 + b^g_1(\text{Post}_d) + b^g_2(\text{Treat}^g_j) + b^g_3(\text{Post}_j) \times (\text{Treat}^g_d) + u_{j,d}. \tag{3}$$

Here $d \in [-11, -1]$ indexes the 11 trading days ending on 10/02/2016, and $d \in [0, 10]$ indexes the 11 trading days beginning on 10/17/2016.[65] $X_{j,d}$ is stock j's outcome variable on trading day d;

[63] Non-midpoint sub-penny trade executions remain available for Group 3 stocks through exchange retail liquidity programs. However, these executions do not involve wholesalers.

[64] Highlighting the binding nature of this constraint for test Group 2 stocks, Figure C.1 illustrates that absent the minimum 0.5¢ PI restrictions, wholesalers offer only 0.01¢ PI most of the time, implying that this restriction raised the PI-driven cost of internalization by a factor of 50 for most internalized trades.

[65] Our event window excludes the 10 trading days spanning 10/03/2016 through 10/16/2016 to account for the staggered phase-in of tick size changes for treated stocks. There were three phase-ins of treated stocks in Test Groups 1 and 2 stocks: 5 stocks from each group on 10/03/2016, 92 stocks from each group on 10/10/2016, and the remaining

Post_d is an indicator variable that equals 0 if $d < 0$ and 1 if $d \geq 0$. Treatment_j^g is an indicator variable that equals 0 if stock j is in the control group and 1 if stock j is in the treatment group for Test Group $g \in \{1, 2\}$. The coefficient b_3^g captures the treatment effects associated with Test Group g. To ensure that estimated treatment effects are unaffected by outliers, we use both OLS and quantile (median) regressions to estimate equation (A.2). Following standard practice (see Rindi and Werner (2019), Griffith, Roseman, and Shang (2020), Albuquerque, Song, and Yao (2020), we condition estimates on quoted spread levels prior to the introduction of TSP.

We obtain the identifying information for control and treatment stocks in the U.S. Tick Size Pilot program (TSP) from FINRA's website, focusing on Test Groups 1 and 2. For each stock, we construct daily observations over the 10 trading days prior to implementation of TSP on 10/03/2016 as well as the 10 trading days after full implementation on 10/17/2016.[66] From Daily TAQ's Trades, Quotes, and NBBO files, we obtain trade and quote information to match off-exchange transactions executed at sub-penny prices with the national best bid and ask prices at the time of transaction based on millisecond timestamps. Then, for each stock-day, we construct the following outcome variables: (1) the absolute value of $Mroibtrd$; (2) the absolute value of $Mroibvol$; (3) size-weighted average relative percentage price improvement, which divides the relative price improvement for a sub-penny-executed transaction (i.e., the difference between the best quoted price and the transaction price) by the mid-point of best bid and ask; (4) total dollar-denominated price improvement, which is the sum of dollar relative price improvements across all sub-penny-executed transactions; (5) the total share volume of trades receiving price improvement; and (6) the size-weighted average sub-tick (sub-penny) fraction of trades receiving price improvement.

Table A.1 presents estimation results for Test Group 1. Panels A-C in Figure A.1 provide complementary visual evidence. The quote rule raises the average and median volume of sub-penny-executed trades by 9% and 63% relative to the corresponding intercept, respectively.[67] This indicates that the quote rule causes wholesalers to internalize retail orders more aggressively. The effects are stronger for stocks with tighter pre-TSP quoted spreads—stocks that are more likely to

303 stocks on 10/17/2016.

[66] Implementation consists of three phase-ins with different subsets of control stocks experiencing tick size changes on 10/03/2016, 10/10/2016, and 10/17/2016. For more details about the Tick Size Pilot program, see https://www.sec.gov/rules/sro/nms/2015/34-74892.pdf.

[67] Rindi and Werner (2019) find no discernible effect on consolidated volumes of treated stocks in TSP, indicating that our findings are likely orthogonal to any stock-level volume effect.

have binding quote test restrictions.

Table A.1. Retail Order Internalization and Tick Size Pilot *Quote Rule*. This table reports OLS and Quantile (median) Regression (QR) estimates of equation (A.2), comparing stocks in Test Group 1 to control stocks. Panels A and C report results for stocks whose average quoted spread in during August, 2016 was below sample median; and Panels B and D report results for stocks with above-median spreads. Sample periods spans the 10 trading day prior to implementation of TSP on 10/03/2016 as well as the 10 trading days following the full implementation of TSP on 10/17/2016 for Test Group 1 stocks. Outcome variables are constructed using trade and quote information of sub-penny-executed off-exchange transactions, and they include (1) the absolute value of $Mroibtrd$; (2) the absolute value of $Mroibvol$; and (3) the total share volume, in round lots, of trades receiving price improvement (PI shr vol). Numbers in brackets reflect t-statistics, and symbols ***, **, and * identify statistical significance at the 1%, 5%, and 10% type one errors, respectively.

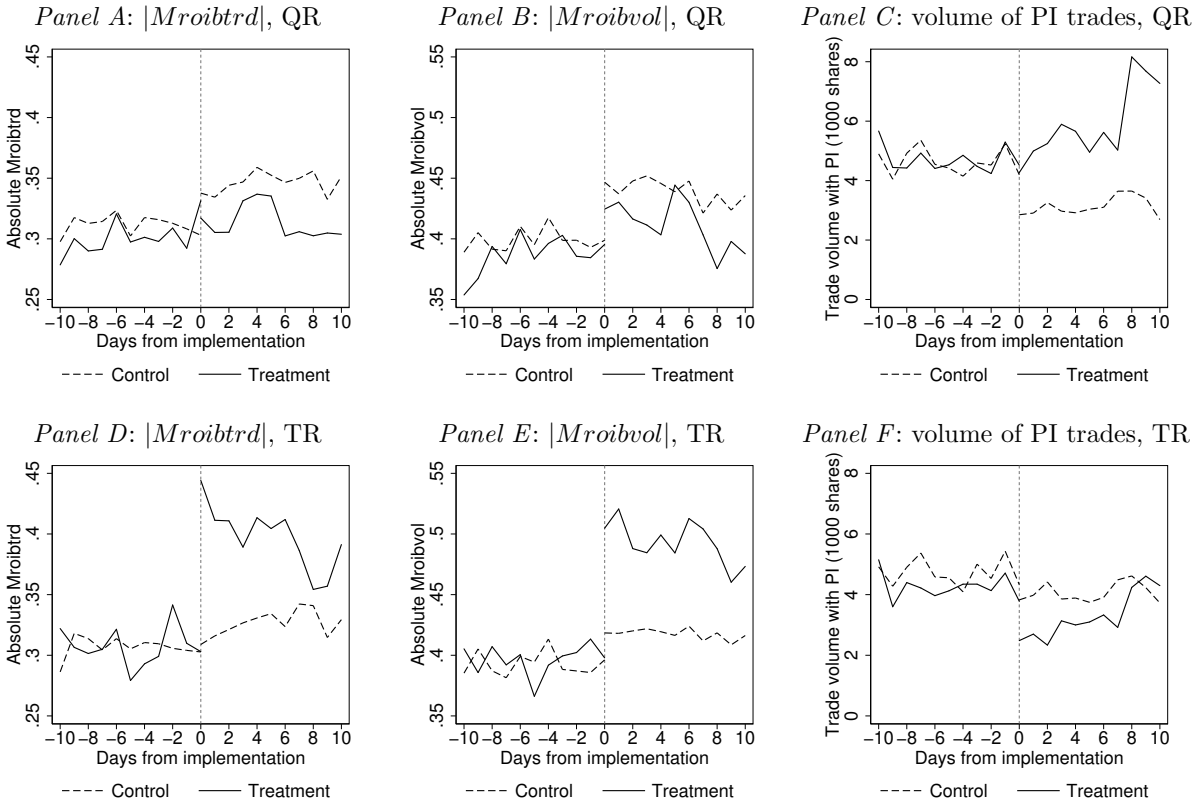
	Panel A: Low-spread stocks, OLS			Panel B: High-spread stocks, OLS		
Outcome variable:	$\lvert Mroibtrd\rvert$	$\lvert Mroibvol\rvert$	PI shr vol	$\lvert Mroibtrd\rvert$	$\lvert Mroibvol\rvert$	PI shr vol
Intercept	0.31***	0.39***	14517.1***	0.31***	0.39***	14517.1***
	[198.21]	[225.90]	[74.77]	[173.36]	[208.29]	[89.14]
PrePost	−0.047***	−0.047***	6277.4***	0.10***	0.12***	−8964.6***
	[−17.74]	[−16.08]	[18.90]	[32.11]	[34.77]	[−32.32]
Treat	−0.012***	−0.0099**	462.3	−0.012***	−0.0099**	462.3
	[−3.15]	[−2.33]	[0.97]	[−2.76]	[−2.15]	[1.16]
PrePost*Treat	0.0034	0.0015	1360.3*	−0.019**	−0.010	−334.1
	[0.54]	[0.21]	[1.70]	[−2.46]	[−1.25]	[−0.49]
	Panel C: Low-spread stocks, QR			Panel D: High-spread stocks, QR		
Outcome variable:	$\lvert Mroibtrd\rvert$	$\lvert Mroibvol\rvert$	PI shr vol	$\lvert Mroibtrd\rvert$	$\lvert Mroibvol\rvert$	PI shr vol
Intercept	0.23***	0.32***	4893***	0.23***	0.32***	4893***
	[132.83]	[136.29]	[71.81]	[102.92]	[112.02]	[107.96]
PrePost	−0.029***	−0.040***	4389***	0.097***	0.14***	−3506***
	[−9.81]	[−10.07]	[37.65]	[24.58]	[27.06]	[−45.42]
Treat	−0.014***	−0.015***	−86	−0.014***	−0.015**	−86
	[−3.40]	[−2.62]	[−0.52]	[−2.63]	[−2.15]	[−0.78]
PrePost*Treat	0.014**	0.011	3057***	−0.023**	−0.0075	927***
	[2.04]	[1.18]	[10.87]	[−2.44]	[−0.61]	[4.85]

Consider a low spread stock for which the 5¢ minimum spread reflects an exogenously-widened quoted spread. For example, suppose marketable limit buy and sell orders were quoted at best prices of $10.02 and $9.99, respectively, before the spread was widened to $10.03 and $9.98. This widening of the spread increases depth at the best price, facilitating larger transactions (Rindi and Werner 2019). However, the aggregate amount of order flow that a wholesaler would otherwise have internalized is unaffected,[68] replacing the set of attractive non-marketable limit orders with marketable limit orders.[69] More importantly, widening the quoted spread increased the profitability of off-exchange liquidity provision at the midpoint, increasing the willingness of wholesalers to

[68] Werner et al. (2019) find that the wider spread incentivized the submission of limit orders, resulting in a longer queue at the bid and ask, while volume was unchanged.

[69] For example, consider two stocks, one with a mandated 5¢ spread and one with a non-mandated (pre-existing) 5¢ spread. There can be attractive non-marketable limit orders with the latter but not the former.

Figure A.1. Tick Size Pilot. This figure provides visual evidence associated with the results of the Difference-in-Difference specification in equation (A.2) for Test Group 1 and Test Group 2. The sample period spans the 10 trading days prior to the TSP's implementation on 10/03/2016 as well as the 10 trading days following its full implementation on 10/17/2016. The figure plots the daily medians for six outcome variables across the control and treatment groups. The outcome variables are constructed using trade and quote information for sub-penny-executed off-exchange transactions and include: the absolute value of $Mroibtrd$; the absolute value of $Mroibvol$; and the total share volume of trades receiving price improvement. Panels A-C and D-F present findings associated with the Quote Rule (QR) and Trade Rule (TR), respectively.



internalize order flow.

Table A.1 reports that the intensity of sub-penny-executed retail trades—as measured by the total volume of price-improved trades—rises due to the minimum 5¢-spread. In contrast, the absolute values of $Mroibvol$ and $Mroibtrd$ fall, moving in the *opposite* direction of retail order flow internalization intensity. That is, $Mroibvol$ and $Mroibtrd$ respond to the economic incentives of wholesalers regarding retail order internalization rather than retail trading per se.

Table A.2 presents estimation results for Test Group 2 that introduced a 0.5¢ minimum PI in addition to the 5¢ pricing increment. Panels D–F in Figure A.1 provide complementary visual evidence. In contrast to the quote-rule treatment, the trade-rule treatment caused the absolute values of $Mroibtrd$ and $Mroibvol$ to increase dramatically, even though the treatment *reduced* the

volume of internalized (sub-penny) trades. For stocks with tight spreads, median internalized trade volume fell by 47%relative to the corresponding intercept, while trade volume is unchanged for stocks with wide spreads.[70]

Table A.2. Retail Order Internalization and Tick Size Pilot *Trade Rule*. This table reports OLS and quantile (median) regression estimates of equation (A.2), comparing stocks in Test Group 2 to control stocks. Panels A and C report results for stocks whose average quoted spread in during August, 2016 was below sample median; and Panels B and D report results for stocks with above-median spreads. Sample periods spans the 10 trading day prior to implementation of TSP on 10/03/2016 as well as the 10 trading days following the full implementation of TSP on 10/17/2016 for Test Group 1 stocks. Outcome variables are constructed using trade and quote information of sub-penny-executed off-exchange transactions, and they include (1) the absolute value of $Mroibtrd$; (2) the absolute value of $Mroibvol$; and (3) the total share volume, in round lots, of trades receiving price improvement (PI shr vol). Numbers in brackets reflect t-statistics, and symbols ***, **, and * identify statistical significance at the 1%, 5%, and 10% type one errors, respectively.

	Panel A: Low-spread stocks, OLS			Panel B: High-spread stocks, OLS		
Outcome variable:	$\lvert Mroibtrd\rvert$	$\lvert Mroibvol\rvert$	PI shr vol	$\lvert Mroibtrd\rvert$	$\lvert Mroibvol\rvert$	PI shr vol
Intercept	0.31***	0.39***	14695.6***	0.31***	0.39***	14695.6***
	[198.89]	[225.93]	[75.76]	[172.60]	[207.06]	[90.92]
PrePost	−0.056***	−0.065***	7917.6***	0.087***	0.10***	−8872.9***
	[−21.80]	[−22.28]	[23.91]	[27.91]	[31.63]	[−32.19]
Treat	0.0043	0.011**	−1382.4***	0.0043	0.011**	−1382.4***
	[1.13]	[2.53]	[−2.92]	[0.98]	[2.32]	[−3.51]
PrePost*Treat	0.032***	0.076***	−3277.9***	0.042***	0.052***	591.6
	[5.13]	[10.79]	[−4.07]	[5.44]	[6.27]	[0.88]
	Panel C: Low-spread stocks, QR			Panel D: High-spread stocks, QR		
Outcome variable:	$\lvert Mroibtrd\rvert$	$\lvert Mroibvol\rvert$	PI shr vol	$\lvert Mroibtrd\rvert$	$\lvert Mroibvol\rvert$	PI shr vol
Intercept	0.22***	0.31***	4948***	0.22***	0.31***	4948***
	[125.61]	[131.66]	[71.84]	[97.95]	[111.11]	[109.61]
PrePost	−0.036***	−0.052***	5796***	0.075***	0.12***	−3296***
	[−11.86]	[−13.06]	[49.29]	[18.57]	[23.75]	[−42.81]
Treat	0.0058	0.0065	−546***	0.0058	0.0065	−546***
	[1.31]	[1.12]	[−3.25]	[1.03]	[0.94]	[−4.96]
PrePost*Treat	0.027***	0.091***	−2326***	0.028***	0.092***	120
	[3.71]	[9.32]	[−8.13]	[2.75]	[7.45]	[0.64]

In Group 2 stocks, the trade rule's minimum 0.5¢ PI requirement sharply raises the costs of internalizing retail orders. The increases in $\lvert Mroibtrd\rvert$ and $\lvert Mroibvol\rvert$ let us attribute the increased variation in $Mroib$ to this increased cost.[71] We posit that these effects manifest themselves in

[70] Our findings are robust to correcting for multiple-testing issues due to reusing natural experiments. Almost all t-statistics associated with the significant treatment effects in Tables A.1 and A.2 exceed the heuristic critical values of 2.5 and 3.0 proposed by Heath et al. (2022).

[71] The increased variation in $Mroib$ may also reflect the increased share of non-marketable limit orders in all internalized order flow. The trade rule quintupled the trading increment. This impacted the composition of retail orders: as market orders risked execution at prices 5¢ further from current best prices (i.e., by more than 1¢), retail traders would rely more on marketable limit orders in lieu of market orders. By the time a wholesaler handles orders flagged as marketable limit, some will have become non-marketable due to updates in the order book, increasing the share of non-marketable limit orders, and hence reducing internalization. Again, internalization is reduced by less when there is (more profitable) institutional demand on the other side than when are retail market orders, resulting

the increased sensitivity of $Mroib$ to institutional liquidity demand, as the orders that are more costly to internalize are the marginal retail orders used to provide liquidity to institutions through internalization. Section C.3 provides further support for this prediction when $Mroib$ is constructed from retail orders with price improvement levels that are relatively more likely to be associated with internalized orders executed at prices falling over 1¢ inside the NBBO.

These findings based on the TSP reinforce conclusions that variations in $Mroibtrd$ and $Mroibvol$ are largely not due to imbalances in the underlying retail order flow. Instead, these measures reflect wholesaler decisions of whether to internalize retail order flow. Our findings also indicate that $Mroib$ is unlikely to capture directional informed retail trading. Interpreting the higher $|Mroib|$ associated with Test group 2 stocks as due to increased informed retail trading would imply that wholesalers pay *more* PFOF + PI to internalize more toxic (informed) retail orders. This is hard to reconcile with any notion of profit-maximization by wholesalers. In contrast, the willingness to pay more for internalizing these marginal orders is consistent with wholesalers facilitating liquidity provision when institutional demand is high. Having established that wholesaler internalization choices are responsible for variation in $Mroib$, we now examine the cross-sectional variation in $Mroib$.

B Signed *Mroib*'s Return Predictability

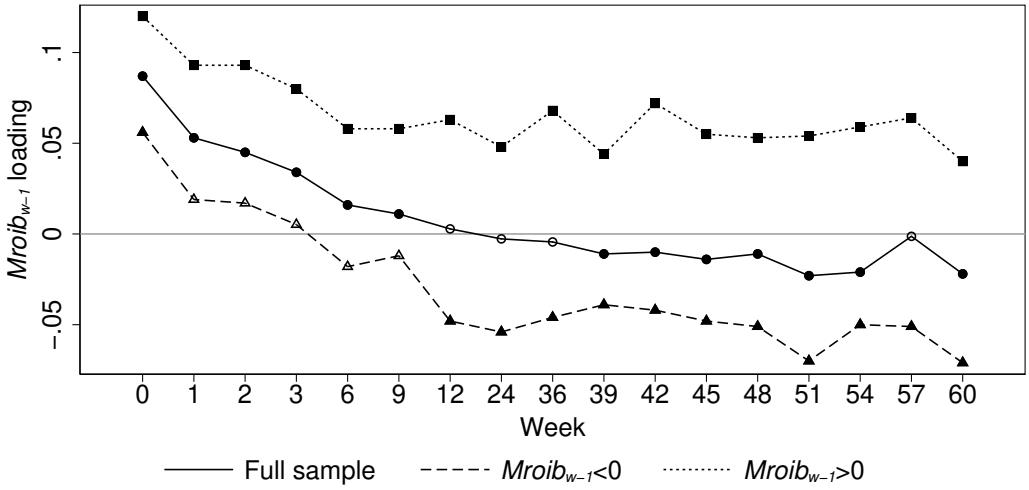
In this section, we examine the return predictability of $Mroib$ in more detail. Our findings are inconsistent with $Mroib$ capturing informed retail order flow. In contrast, near-term future weekly returns conditional on $Mroib$ are consistent with price reversals following liquidity consumption by institutional investors.

We estimate (1) both unconditionally and conditional on the *sign* of $Mroibvol_{j,w-1}$ to examine its return predictability separately when this order flow imbalance is negative and positive. As in BJZZ, we estimate equation (1) using Fama-Macbeth regressions, featuring Newey-West corrected standard errors with 6 lags. We extend their analysis in three ways. First, we estimate the weekly return predictability of $Mroibvol_{j,w-1}$ for up to 60 weeks ahead (past the 12 weeks in BJZZ). Second, we estimate return predictability conditional on the sign of $Mroibvol_{j,w-1}$. Third, we decompose returns entering the left-hand-side of equation (1) into intraday and overnight components.

Striking evidence obtains. Figure B.1 shows that the coefficients on $Mroibvol_{j,w-1}$ become

in more unbalanced $Mroib$.

Figure B.1. Internalized Order Flow and the Cross-sections of Future Weekly Returns. This figure shows the associations between $Mroibvol_{w-1}$ and future week $w+i$ returns (in %), with $i \in \{0, 1, 2, 3, 6, 9, 12, 24, 36, 39, 42, 45, 48, 51, 54, 57, 60\}$. Returns reflect the quoted mid-points at the close. According to equation (1), week $w+i$ returns in each sample are regressed on $Mroibvol_{w-1}$, whose loadings are plotted in future weeks for both the unconditional analysis and the analysis conditional on the sign of $Mroibvol_{w-1}$. The estimated loadings are from Fama-Macbeth regressions, featuring Newey-West standard errors with 6 lags. Statistically significant and insignificant $Mroibvol_{w-1}$ loadings at the 10% type one error are identified by *filled* and *hollow* markers, respectively. The sample includes NMS common shares from January 2010 to December 2014, excluding observations when the previous month-end's closing price is below $1.



uniformly negative after 39 weeks. This is inconsistent with informed retail trading, but consistent with return dynamics reflecting pricing errors (Hendershott, Menkveld, Praz, and Seasholes (2022)). The far-future return reversals are also consistent with the positive association between $Mroib$ and changes in short interest documented in Table 2. As established by the literature, increased short interest (associated with higher $Mroib$) predicts lower future returns, while decreased short interest (associated with lower $Mroib$) predicts higher future returns (Desai et al. (2002); Engelberg et al. (2012); Boehmer and Wu (2013)). Moreover, although a negative $Mroibvol_{j,w-1}$ yields a positive coefficient for the current week's close-to-close return ($i = 0$), this coefficient declines and becomes *negative* by week $w+6$, contrary to retail sell orders being informed, as "retail sell order flow" realizes weekly losses due to persistent price appreciation after 6 weeks. In contrast, a positive $Mroibvol_{j,w-1}$ yields a positive coefficient for weekly returns across all horizons.

Decomposing returns into intraday and overnight components uncovers further asymmetries in the loadings conditional on the sign of $Mroibvol_{j,w-1}$. For overnight returns, $\hat{c}_w^1$ is positive after negative $Mroibvol_{j,w-1}$ (retail selling, institutional buying), but negative and insignificant after positive $Mroibvol_{j,w-1}$ (retail buying, institutional selling). Barclay and Hendershott (2003) and

Jiang, Likitapiwat, and T. McInish (2012) show that overnight price movements are information-driven; the insignificant negative relation between net retail buying imbalances and next week's overnight returns indicates that retail buys are not informed.[72] Moreover, informed retail trading cannot explain why $\hat{c}_w^1$ switches sign for intraday returns when $Mroibvol_{j,w-1}$ switches sign.[73]

C Why Does *Mroib* Predict Short-Term Returns?

In this section, we report how wholesaler liquidity provision to institutional investors is responsible for the return predictability of $Mroib$. Specifically, we attribute this return predictability to the unwinding of institutional price pressure.

C.1 Dynamics of Institutional and Retail Order Flows

In Section 5.2, we documented that overnight reversals exceeded intraday price pressure (in the same week). This section reconciles this phenomenon by showing that overnight reversals also reflect the unwinding of institutional price pressure accumulated in prior weeks. This effect is more salient when more retail sell orders have been internalized, presumably to provide liquidity for institutional buy orders.

To show this, we estimate

$$\begin{aligned} X_{j,w} &= a^0 + \sum_{i=1}^{6} a_i^1 Inoibvol_{j,w-i} + \sum_{i=1}^{6} a_i^2 \left[I(Inoibvol_{j,w-i} < 0)\right] \\ &+ \sum_{i=1}^{6} a_i^3 \left[I(Inoibvol_{j,w-i} < 0) \times Inoibvol_{j,w-i}\right] + \epsilon_{j,w}, \end{aligned} \tag{4}$$

where $X \in \{Inoibvol, Mroibvol\}$; and $I(\cdot)$ is an indicator function that equals 1 if $Inoibvol < 0$ and equals 0 otherwise. The models are estimated using Fama-MacBeth regressions, with standard errors corrected using the Newey-West methodology with 6 lags. On average across stocks, ANcerno covers less than 7% of the total daily trading volume reported by CRSP.[74] To reduce the noise attributable to a lack of coverage we use the subset of stocks for which the share of ANcerno-

[72] Furthermore, retail short selling is limited, suggesting that informed trading does not underlie the association between net retail selling imbalances and next week's overnight returns.

[73] Table **??** shows that the asymmetry in the predictability of close-to-close returns also holds for intraday and overnight returns, which is further at odds with retail investors being informed.

[74] Hu et al. (2018) report similar coverage over a longer sample period. However, modest coverage does not invalidate the representativeness of ANcerno data (Puckett and Yan 2011; Anand et al. 2012; Jame 2018).

reported volume relative to CRSP is above-average.

Columns (1)–(4) in Table C.1 present the AR(k) estimates for $Inoibvol$, showing that past positive and negative institutional trade imbalances, especially those for institutional buying, predict current institutional trade imbalances differently. The most recent week's positive and negative $Inoibvol$ predict current week's $Inoibvol$ similarly, with point estimates of 0.33 and 0.35 for positive and negative $Inoibvol_{w-1}$, respectively. However, these coefficients sharply diverge for $k > 1$, where the loadings of negative $Inoibvol_{w-i}$ become 30-70% smaller than those on their positive $Inoibvol_{w-i}$ counterparts. This finding is consistent with a literature that finds long-only fund managers accumulate long positions slowly, but sell quickly, largely to fund purchases.[75] This persistent institutional buying drives the accumulation of positive price pressure whose unwinding extends beyond the subsequent close-to-open to subsequent days, while institutional selling is less persistent.

Columns (5)–(8) in Table C.1 highlight how past institutional trade imbalances predict future internalized retail order flow, reinforcing our earlier conclusion that wholesalers intermediate trades between institutional and retail investors. Consistent with the stronger auto-correlation for institutional buying, and retail sell orders being internalized to provide liquidity for institutional buy orders, $Inoibvol_{w-i}$ loads with negative and significant coefficients.[76] Mirroring the weaker auto-correlation in institutional trade imbalances when $Inoibvol_{w-i} < 0$, the loadings for $Inoibvol_{w-i}$ become positive for $k > 2$. These dynamics indicate that the most negative $Mroibvol_w$ observations, i.e., those in decile 1 of Table 2, are disproportionately more likely to arise following persistent institutional buying pressure whose unwinding makes the current week's overnight returns more negative.

These statistical findings contain insights about institutions' demand for retail sourced liquidity. The negative correlation between past positive institutional trade imbalances and current internalized retail order flow is consistent with institutions resorting to retail-sourced liquidity, provided by wholesalers, especially in less liquid markets.

[75] This asymmetry is consistent with institutional buying, but not selling, being motivated by a fund manager's best ideas (Akepanidtaworn et al. 2021). This leads managers to accumulate long positions more slowly to conceal their presence, prolonging the unwinding of price pressure. Hendershott and Seasholes (1994) also find that short positions of market makers, which are accumulated due to institutional buying, are associated with subsequent price reversals that last up to 11 trading days. In contrast, price reversals following the accumulation of long positions by market markers, which reflect institutional selling, only last for 7 trading days.

[76] The only exception to statistical significance appears in column (8) for $Inoibvol_{w-5}$.

Table C.1. Asymmetric Persistence in Institutional Trade Imbalances: Implications for Retail Flow Internalization. This table presents estimates of the predictive power of past institutional trade imbalance, conditional on its sign, for both current institutional trade imbalance and current internalized retail order flow. Columns (1)–(4) report estimation results of equation (4) for $i \in \{3, 4, 5, 6\}$ and $X = Inoibvol_w$. Columns (5)–(8) report estimation results of equation (4) for $i \in \{3, 4, 5, 6\}$ and $X = Mroibvol_w$. Fama-MacBeth regressions are used with Newey-West-corrected standard errors using 6 lags. The sample contains stocks with average ANcerno-to-CRSP daily volume of 6.8% or higher. Numbers in brackets reflect t-statistics, and symbols ***, **, and * identify statistical significance at the 1%, 5%, and 10% type one errors, respectively.

	Dependent variable: $Inoibvol_w$				Dependent variable: $Mroibvol_w$			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.065**	0.038	0.023	0.0088	−0.16***	−0.15***	−0.15***	−0.14***
	[2.46]	[1.42]	[0.86]	[0.32]	[−14.54]	[−13.32]	[−12.50]	[−11.77]
$Inoibvol_{w-1}$	0.33***	0.33***	0.33***	0.33***	−0.016***	−0.016***	−0.016***	−0.016***
	[58.36]	[59.42]	[58.50]	[58.00]	[−7.43]	[−7.86]	[−7.66]	[−7.58]
$I(Inoibvol_{w-1} < 0) \times Inoibvol_{w-1}$	0.020***	0.022***	0.022***	0.023***	0.0083***	0.0085***	0.0081**	0.0085***
	[2.71]	[2.98]	[3.03]	[3.24]	[2.63]	[2.69]	[2.56]	[2.65]
$Inoibvol_{w-2}$	0.075***	0.072***	0.071***	0.069***	−0.0067***	−0.0062***	−0.0060***	−0.0059***
	[17.07]	[16.60]	[16.60]	[15.35]	[−3.41]	[−3.06]	[−2.94]	[−2.90]
$I(Inoibvol_{w-2} < 0) \times Inoibvol_{w-2}$	−0.023***	−0.020***	−0.020***	−0.018**	0.0059*	0.0051	0.0046	0.0044
	[−3.06]	[−2.70]	[−2.68]	[−2.46]	[1.85]	[1.57]	[1.38]	[1.31]
$Inoibvol_{w-3}$	0.062***	0.048***	0.045***	0.043***	−0.0069***	−0.0054***	−0.0052**	−0.0050**
	[13.26]	[10.52]	[9.90]	[9.65]	[−3.40]	[−2.64]	[−2.53]	[−2.41]
$I(Inoibvol_{w-3} < 0) \times Inoibvol_{w-3}$	−0.017***	−0.014**	−0.012*	−0.011*	0.0091***	0.0079***	0.0078***	0.0077**
	[−2.63]	[−2.14]	[−1.86]	[−1.79]	[3.09]	[2.66]	[2.63]	[2.54]
$Inoibvol_{w-4}$		0.052***	0.040***	0.037***		−0.0055***	−0.0048**	−0.0050**
		[12.29]	[9.65]	[8.77]		[−2.64]	[−2.30]	[−2.40]
$I(Inoibvol_{w-4} < 0) \times Inoibvol_{w-4}$		−0.023***	−0.021***	−0.019***		0.0078**	0.0080***	0.0078***
		[−3.51]	[−3.20]	[−2.90]		[2.58]	[2.69]	[2.60]
$Inoibvol_{w-5}$			0.041***	0.031***			−0.0041**	−0.0028
			[10.22]	[7.73]			[−2.11]	[−1.38]
$I(Inoibvol_{w-5} < 0) \times Inoibvol_{w-5}$			−0.029***	−0.025***			0.00047	0.000084
			[−4.14]	[−3.78]			[0.16]	[0.03]
$Inoibvol_{w-6}$				0.037***				−0.0044**
				[9.35]				[−2.15]
$I(Inoibvol_{w-6} < 0) \times Inoibvol_{w-6}$				−0.026***				0.0019
				[−3.79]				[0.63]
Observations	976,110	976,110	976,110	976,110	976,110	976,110	976,110	976,110

C.2 Institutional Trading and Short-Term Return Predictability

We next establish that $Mroib$'s short-term return predictability is a liquidity-driven phenomenon. Due to the persistence of institutional liquidity demand, especially institutional buying, overnight price reversals associated with extreme $Mroibvol$ magnitudes extend into future weeks. This creates distinguishable differences between close-to-close returns that follow extremely negative and extremely positive internalized retail order flow imbalances.

To highlight the persistence of institutional liquidity demand, we estimate

$$\begin{aligned} Inoibvol_{j,w} &= c^0 + \sum_{i=1}^{6} c_i^1 Mroibvol_{j,w-i} + \sum_{i=1}^{6} c_i^2 \left[I(Inoibvol_{j,w-i} < 0)\right] \\ &+ \sum_{i=1}^{6} c_i^3 \left[I(Inoibvol_{j,w-i} < 0) \times Mroibvol_{j,w-i}\right] + \epsilon_{j,w}. \end{aligned} \tag{5}$$

Variable definitions and estimation approaches are identical to those in equation (4). Table C.2 shows that the first and second lags of internalized retail order flow load with significantly negative coefficients when these lagged internalized order flows correspond to positive institutional flow. That is, when institutional buy pressure is higher, the greater internalization of retail sell orders relative to buy orders is associated with abnormally high institutional buy pressure for up to two weeks ahead. This persistence drives subsequent abnormally negative overnight returns, due to reversals after institutional price pressure that skew future weeks' close-to-close returns downward. Thus, while *Mroibvol* seems to predict future close-to-close returns, this just reflects price reversals following institutional buy pressure.

C.3 Implications of the Size of Price Improvement

To provide further support for how wholesaler choices drive *Mroib* imbalances, we now delve more deeply into the link between institutional liquidity demand and the magnitudes of sub-penny price improvements that wholesalers offer when internalizing retail orders. We show that stronger institutional demand for liquidity, as manifested by more extreme institutional trade imbalance and price impacts, is associated with more costly internalization, i.e., internalized retail orders not only with larger sub-penny price improvements but also a higher probability of execution at prices inside the NBBO by over 1¢.

Figure C.1 plots the histogram of sub-penny price improvements associated with internalized retail trades, as identified by BJZZ's algorithm. Over 80% of sub-penny PIs are at 0.01¢, 0.1¢, 0.2¢, 0.25¢, 0.3¢, or 0.4¢ increments, suggesting that simple informal agreements govern price improvement schedules. More importantly, we find that (1) the size of price improvement is positively related to the bid-ask spread; (2) the sub-penny increments of PIs are larger when internalized orders are executed inside the NBBO by over 1¢; and (3) more frequent such inside-quote internalization is associated with wider bid-ask spreads.

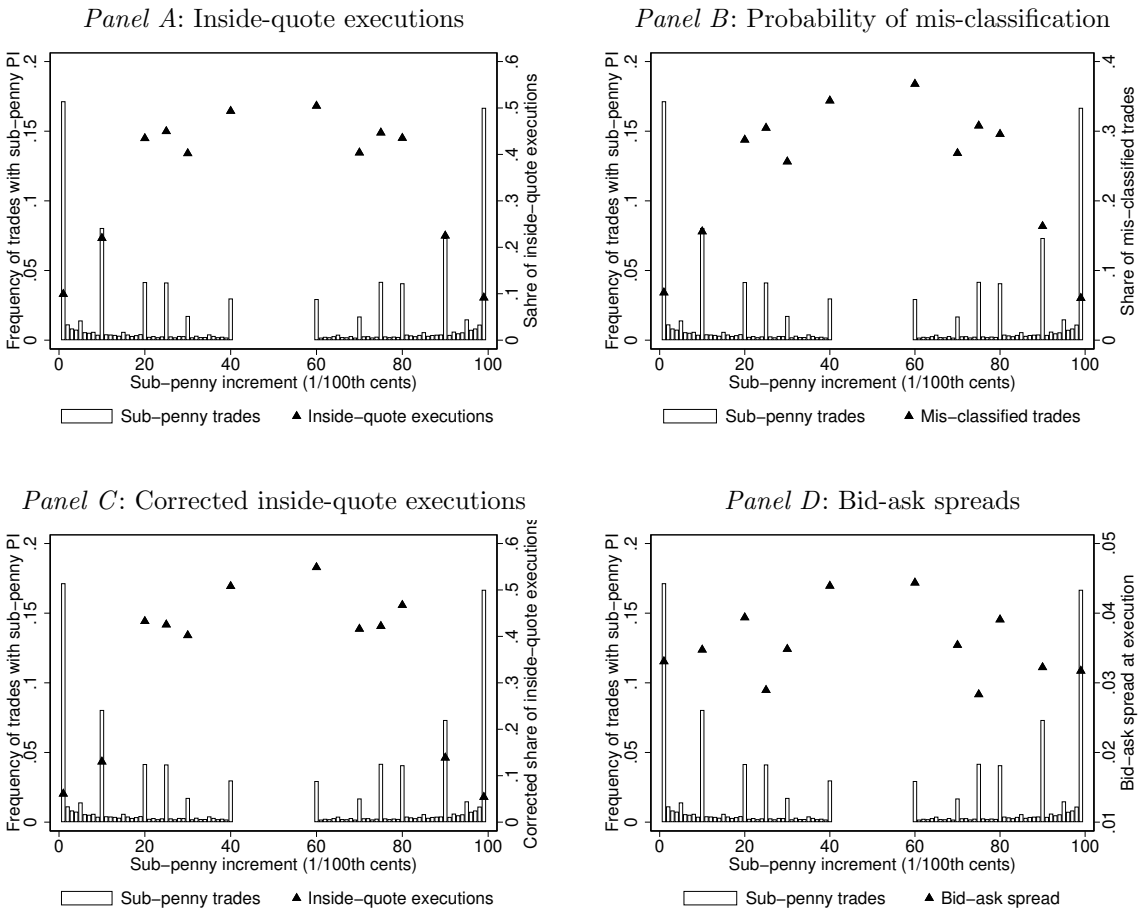
Table C.2. Predictability of Institutional Trade Imbalances Using Internalized Retail Trading Imbalance. This table presents estimates of the predictive power of past internalized order flow, conditional the sign the corresponding institutional trade imbalance, for current institutional trade imbalance. Equation (5) for $i \in \{3, 4, 5, 6\}$ and $X = Inoibvol_w$ is estimated using Fama-MacBeth regressions with Newey-West-corrected standard errors using 6 lags. The sample contains stocks with average ANcerno-to-CRSP daily volume of 6.8% or higher. Numbers in brackets reflect t-statistics, and symbols ***, **, and * identify statistical significance at the 1%, 5%, and 10% type one errors, respectively.

	(1)	(2)	(3)	(4)
Constant	1.04***	1.09***	1.14***	1.17***
	[39.71]	[40.99]	[41.98]	[43.14]
$Mroibvol_{w-1}$	−0.020***	−0.021***	−0.021***	−0.020***
	[−3.69]	[−3.74]	[−3.74]	[−3.57]
$I(Inoibvol_{w-1} < 0) \times Mroibvol_{w-1}$	0.021***	0.020***	0.020***	0.021***
	[2.85]	[2.78]	[2.81]	[2.79]
$Mroibvol_{w-2}$	−0.013**	−0.014**	−0.013**	−0.013**
	[−2.43]	[−2.56]	[−2.43]	[−2.38]
$I(Inroibvol_{w-2} < 0) \times Mroibvol_{w-2}$	0.025***	0.025***	0.025***	0.024***
	[3.41]	[3.39]	[3.42]	[3.30]
$Mroibvol_{w-3}$	−0.0043	−0.0063	−0.0054	−0.0067
	[−0.72]	[−1.13]	[−0.93]	[−1.14]
$I(Inoibvol_{w-3} < 0) \times Mroibvol_{w-3}$	0.017**	0.018***	0.019***	0.020***
	[2.38]	[2.59]	[2.59]	[2.72]
$Mroibvol_{w-4}$		0.0047	0.0054	0.0035
		[0.70]	[0.87]	[0.57]
$I(Inoibvol_{w-4} < 0) \times Mroibvol_{w-4}$		0.0017	0.0038	0.0038
		[0.23]	[0.51]	[0.52]
$Mroibvol_{w-5}$			−0.0058	−0.0065
			[−1.08]	[−1.20]
$I(Inoibvol_{w-5} < 0) \times Mroibvol_{w-5}$			−0.0036	−0.0018
			[−0.45]	[−0.22]
$Mroibvol_{w-6}$				0.0025
				[0.42]
$I(Inoibvol_{w-6} < 0) \times Mroibvol_{w-6}$				0.0056
				[0.63]
Observations	976,110	976,110	976,110	976,110

We first observe that BJZZ's algorithm, which does not require the use of quote data, incorrectly signs some buy retail trades as sells, and vice versa. We describe the source of mis-classification with an example: suppose the NBB and NBO are \$9.97 and \$10.03, and a marketable buy order placed at \$10.03 is executed at \$10.013. BJZZ's algorithm observes the sub-penny increment of 0.3¢ and signs this transaction as a sell, but the trade is actually a buy receiving price improvement of 1.7¢. As Section 5.2 notes, Battalio et al. (2022) show that many trade mis-classifications reflect the algorithm's inclusion of some institutional trades.

Matching each transaction with the corresponding NBBO and comparing execution prices against quote midpoints yields estimates for the share of incorrectly-signed trades by sub-penny

Figure C.1. Distributions of Sub-penny trades, Bid-Ask Spreads, and the Probability of Inside-Quotes Execution. This figure plots a histogram of sub-penny price improvements (in $1/100^{th}$ cents) associated with transactions. For each stock-year, the frequency of trades associated with each of the 80 sub-penny increments, from 0.01¢ through 0.40¢ and from 0.60¢ through 0.99¢ is calculated. The mean frequency for a given increment, measured on the left axes of the for panels, is then averaged across stocks and years. The figure also reports, for the 12 most frequent sub-penny price improvement outcomes, (1) the share of corresponding transactions executed by at least 1¢ inside the NBBO (Panel A); (2) the share of transactions mis-classified by the BJZZ algorithm (Panel B); (3) the share of corresponding transactions executed by at least 1¢inside the NBBO after removing mis-classified trades from the sample (Panel C); and (4) average bid-ask spread at the time of execution (Panel D).



increment.[77] Panel B in Figure C.1 shows that this share rises sharply with the distance of the sub-penny increment from the nearest full penny. Importantly, an unreported robustness analysis reveals that all of our main findings continue to hold when we correct for the mis-classification of trades, likely because our aggregation to the weekly level mitigates the largely idiosyncratic nature of mis-classified buy and sell trades. Complementing our findings in Section 5.2, this robustness finding indicates that imbalances in sub-penny executed institutional trades not reported

[77] Barber et al. (2022) use a similar approach to identify signing errors in BJZZ's algorithm.

by ANcerno do not drive the variation in $Mroib$.

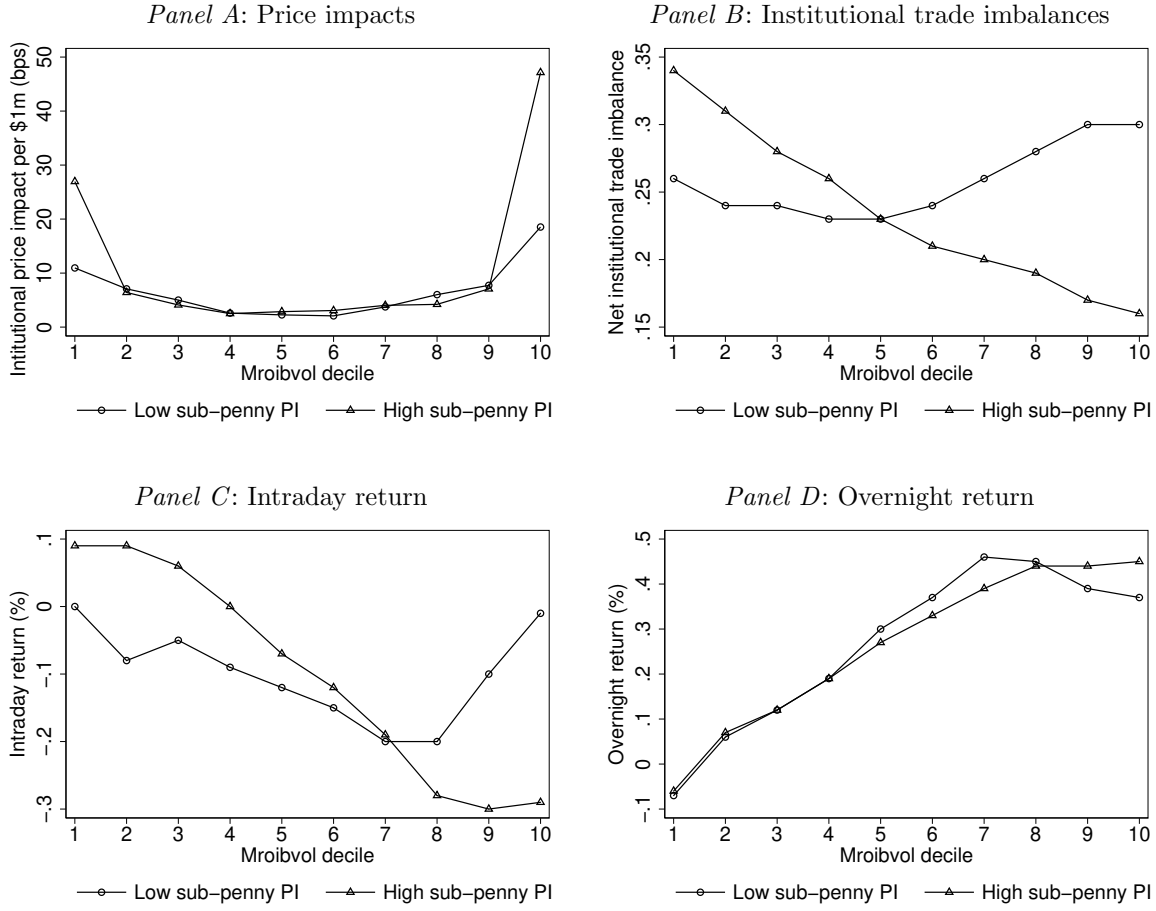
We next document a strong relation between the size of sub-penny increment of PIs and the likelihood that the corresponding execution takes place inside the NBBO by over 1¢. This finding is robust to removing mis-classified retail trades from the sample. As Panels A and C in Figure C.1 illustrate, the share of internalized retail orders whose execution price is at least 1¢ better than the NBBO at the time of transaction rises sharply with the size of sub-penny PI increment. For example, after removing mis-classified trades, this share goes from about 5% to over 50% as the sub-penny increment of PI goes from 0.01¢ (0.99¢) to 0.4¢ (0.6¢). Moreover, Panel D of Figure C.1 shows that both larger sub-penny PI increments and more frequent inside-quote executions are associated with wider bid-ask spreads. Overall, our findings suggest that wholesalers are willing to spend more PFOF+PI to internalize orders in less liquid markets. We next show that the imbalance in these more costly internalized orders is more strongly related to institutional trading costs than the imbalance in the less costly internalized orders, highlighting the economic motives that justify such costlier internalization.

Figure C.1 shows that the median sub-penny price improvement is 0.1¢. This leads us to construct two versions of $Mroibvol$, one for internalized retail orders with "small" sub-penny PI increments of less than 0.1¢ and one for "large" such increments of at least 0.1¢.[78] We then compare institutional trading outcomes, price impacts, institutional trade imbalances, intraday returns (proxy for institutional price pressure), and overnight return (proxy for the unwinding of institutional price pressure), across the two versions of $Mroibvol$.

Panel A in Figure C.2 shows that price impacts display far stronger ∪-shaped patterns for high-sub-penny $Mroibvol$ than for low-sub-penny $Mroibvol$. That is, the most extreme high-sub-penny $Mroibvol$ observations occur when institutional trading costs are highest. This result reinforces that the unbalanced internalization of retail orders that are more costly to internalize, due to large price improvements, occurs when wholesalers provide liquidity to institutions willing to incur larger price impacts to locate liquidity. Panel B provides further evidence of this mechanism, showing a sharp inverse relationship between institutional trade imbalances and high-sub-penny $Mroibvol$,

[78] Unreported results establish that the predictive power of $Mroib$ for short-term future returns is not affected by the size of sub-penny PI used to construct $Mroib$ with a 0.1¢ threshold. BJZZ classify transactions into those with small versus large price improvement using a 0.2¢ cutoff. The 0.2¢ threshold assigns over 75% of internalized retail trades to the "small" sub-penny group, resulting in a noisy $Mroib$ based on "large" PI.

Figure C.2. Price Impacts, Institutional Trade Imbalances, Intraday Returns, and Overnight Returns Conditional on the Magnitude of Price Improvement. This figure compares contemporaneous institutional price impacts, institutional net trade imbalance, intraday returns, and overnight returns when $Mroirbvol$ is constructed using retail trades with sub-penny price improvements that are low ($<$ .01¢) versus high ($\geq$.01¢). Stocks are first sorted each day into deciles of low-sub-penny $Mroibvol$ and high-sub-penny $Mroibvol$. Then, each outcome variable is plotted across the deciles of both $Mroibvol$ measures. Panel A plots median price impacts (in basis points per million dollars), Panel B plots average net institutional trade imbalance, Panel C plots average intraday returns, and Panel D plots average overnight returns.



highlighting how institutional liquidity demand drives the unbalanced and costly internalization of retail orders on the opposite side. In contrast, institutional trade imbalance is weakly ∪-shaped conditional on low-sub-penny $Mroibvol$. Building on these insights, Panels C and D show that as high-sub-penny $Mroibvol$ rises, intraday returns fall from 10bps to −30bps while overnight returns reverse in the opposite direction. That is, high-sub-penny $Mroibvol$ is associated with institutional price pressure followed by overnight reversals. In contrast, with small-sub-penny $Mroibvol$, returns mirror the weak ∪-shaped pattern in institutional trade imbalances.

D *ILM*s, Existing Liquidity Measures, and Institutional Price Impacts: Excluding Low Sub-Penny Volume Stocks

This section establishes that the findings documents by Figure 4 and Table 5 are not driven by stocks with low levels of trading volumes executed at sub-penny prices.

Table D.1. Institutional Liquidity Measures and Stock Characteristics. The table reports on the cross-sectional relation between ILMs and (1) three-factor Fama-French betas, (2) book-to-market ratios (BM), (3) natural log of market capitalizations (ln(Mcap)), (4) dividend yields (DYD), (5) idiosyncratic volatilities (IdVol), (6) previous month's returns ($RET_{(-1)}$), and (7) preceding returns from the prior 11 months ($RET_{(-12,-2)}$). Stock characteristics are computed from the prior month. Each weekly cross-section is sorted into ILM deciles. The average outcome variable is calculated by $ILMT$ decile in each cross-section before the average of the time-series is calculated. Panels A and B report the results for $ILMT$ and $ILMV$, respectively. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$2 and stocks falling in the bottom 10% of the share of sub-penny executed volume in total volume.

Panel A: Trade-based Institutional Liquidity Measures ($ILMT$s) versus stock characteristics

	Weekly $ILMT$ deciles									
	1	2	3	4	5	6	7	8	9	10
Stock Characteristics:										
β^{mkt}	1.02	1.02	1.02	1.01	1.00	0.99	0.97	0.93	0.88	0.82
β^{hml}	0.73	0.73	0.73	0.73	0.74	0.75	0.76	0.77	0.78	0.79
β^{smb}	0.15	0.15	0.16	0.16	0.17	0.17	0.18	0.20	0.22	0.24
BM	0.64	0.64	0.65	0.65	0.66	0.67	0.68	0.72	0.76	0.80
ln(Mcap)	20.99	20.98	20.95	20.91	20.85	20.76	20.64	20.38	20.05	19.71
DYD	0.015	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.015	0.015
Id. Vol.	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.022	0.022
$RET_{(-1)}$	0.016	0.018	0.016	0.017	0.016	0.015	0.014	0.015	0.015	0.016
$RET_{(-12,-2)}$	0.19	0.19	0.19	0.19	0.19	0.18	0.17	0.16	0.15	0.14

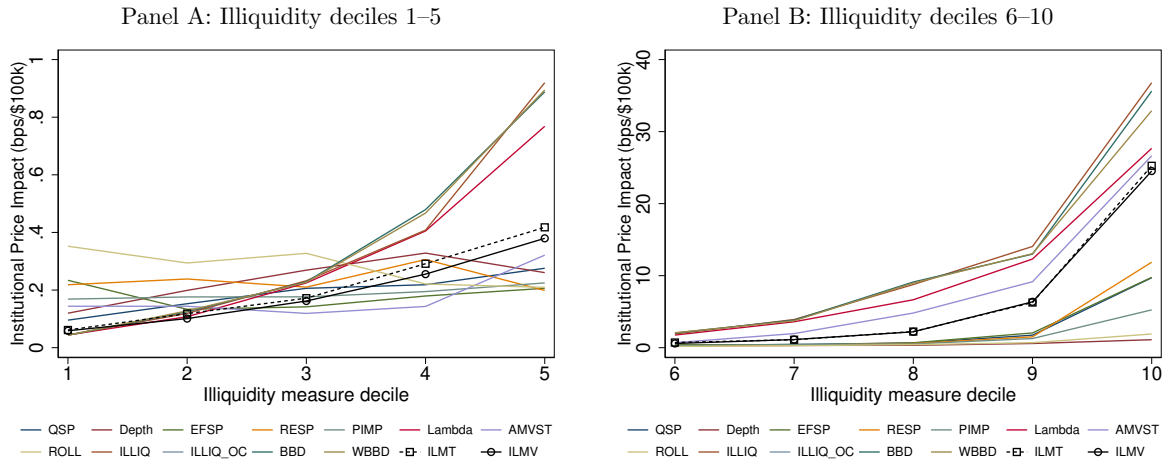
Panel B: Volume-based Institutional Liquidity Measures ($ILMT$s) versus stock characteristics

	Weekly $ILMV$ deciles									
	1	2	3	4	5	6	7	8	9	10
Stock Characteristics:										
β^{mkt}	1.07	1.07	1.06	1.04	1.02	1.00	0.94	0.94	0.89	0.73
β^{hml}	0.71	0.71	0.72	0.73	0.73	0.75	0.74	0.79	0.82	0.77
β^{smb}	0.12	0.12	0.13	0.14	0.15	0.17	0.19	0.21	0.25	0.29
BM	0.62	0.62	0.63	0.63	0.64	0.65	0.70	0.70	0.74	0.87
ln(Mcap)	21.29	21.26	21.19	21.10	20.97	20.81	20.45	20.36	20.01	19.26
DYD	0.015	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.015	0.015
Id. Vol.	0.022	0.022	0.022	0.021	0.021	0.021	0.021	0.020	0.021	0.021
$RET_{(-1)}$	0.019	0.018	0.017	0.016	0.016	0.015	0.014	0.014	0.014	0.015
$RET_{(-12,-2)}$	0.21	0.21	0.20	0.19	0.19	0.18	0.16	0.16	0.15	0.13

A potential concern with ILMs is that these measures do not account for the intensity of trade execution at sub-penny prices, allowing the effects of low sub-penny volume to be conflated with high imbalances in internalized retail orders. For example, suppose that 100,000 shares of both stocks A and B are traded on a given trading day. Also suppose that while stock A, on the same day, has 1,500 shares of buy retail trades and 1,000 shares of sell retail trades executed at sub-penny

prices; stock B has 15,000 shares of buy and 10,000 shares of sell retail trades. For both stocks, $|Mroibvol| = 0.2$, even though retail trading in stock B is far higher than that in stock A. This leads us to examine the robustness of our results to excluding stocks whose share of sub-penny executed volume relative to total trading volume (SPVS) is low. Specifically, Table D.1 and Figure D.1 how that excluding stocks whose SPVS fall in the bottom 10% of each cross-section leaves our qualitative findings unaffected.

Figure D.1. ***ILM*s, Standard Liquidity Measures, and Future Institutional Price Impacts.** The table reports on the cross-sectional relation between various liquidity measures constructed in month $m-2$ and realized, post-trade institutional price impacts, InPrIm, (in bps per \$100k) constructed in month m. Liquidity measures include (1) quoted bid-ask spread (QSP); (2) quoted depth at best prices (Depth); (3) effective spreads (EFSP); (4) realized spreads (RESP); (5) price impacts (PIMP); (6) Kyle's lambda estimates (Lambda); (7) Amvist illiquidity measure (AMVST); (8) Roll measure of realized spreads (ROLL); (9 & 10) close-to-close and open-to-close Amihud measures (ILLIQ & ILLIQ_OC); (11 & 12) simple and volume-weighted trade-time liquidity measures (BBD & WBBD); (13 & 14) trade- and volume-based institutional liquidity measures (ILMT & ILMV). Each month, stocks are sorted into deciles of liquidity, with decile 1 (10) reflecting the most (least) liquid stocks, based on a given liquidity measure from month $m-2$. Month m InPrIm of the median stock in each liquidity decile is averaged across months by liquidity decile. This average is plotted against the respective liquidity decile. Panels A and B report results for liquidity deciles 1 through 5 and 6 through 10, respectively. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$2 and stocks falling in the bottom 10% of the share of sub-penny executed volume in total volume.



E Liquidity and Expected Returns: \$1 and \$5 Share Price Requirements

This section presents estimation results for equation (2) when low-priced stocks are excluded from the sample based on alternative cutoffs for prior month's share prices.

Panel A in Tables E.1 and E.2 reports estimation results when liquidity measures are constructed over one month using samples of stocks with previous month's minimum closing prices of \$1 and \$5, respectively. According to Table E.1, in a more inclusive sample with a less strict (under

Table E.1. Liquidity and the Cross-Section of Expected Stock Returns: 1-month ILMs. This table reports on the relation between alternative high-frequency liquidity measures and the cross-section of expected returns. In Panel A, equation (2) is estimated using liquidity measures ($LIQ_{j,m-2}$) constructed over 1-month horizons. Control variables include three-factor Fama-French betas ($\beta^{mkt}_{j,m-1}$, $\beta^{hml}_{j,m-1}$, $\beta^{smb}_{j,m-1}$), estimated using weekly observations from the two-year period ending in the final full week of month $m-1$, book-to-market ratio, ($\text{BM}_{j,m-1}$), natural log of market capitalization, ($\ln(\text{Mcap}_{j,m-1})$), dividend yield ($\text{DYD}_{j,m-1}$), defined as total dividends over the past 12 months divided by the share price at the end of month $m-1$, idiosyncratic volatility ($\text{IdVol}_{j,m-1}$), previous month's return ($RET_{(-1)}$), and preceding return from the prior 11 months ($RET_{(-12,-2)}$). Panel B replaces each high-frequency liquidity measure by the residuals of $ILMT$ and $ILMV$ with respect to each alternative liquidity measure, with residuals calculated separately for each monthly cross-section. Estimates are from Fama-MacBeth regressions that have Newey-West corrected standard errors with 6 lags. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$1. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Stock liquidity and the cross-section of expected returns

	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD	ILMT	ILMV
Constant	2.03	0.93	0.92	0.90	0.93	0.93	0.80	0.91	1.33	0.88	0.56	1.42	1.26	−0.87	−1.65
	[1.42]	[0.89]	[0.92]	[0.86]	[0.93]	[0.93]	[0.83]	[0.92]	[1.36]	[0.88]	[0.55]	[1.41]	[1.23]	[−0.58]	[−1.03]
Liquidity	**0.024**	**−0.023**	**−0.0000065**	**0.081**	**0.025**	**−0.068**	**0.034**	**0.10**	**−7.04*** **	**0.018**	**0.13** **	**0.18***	**0.39** **	**1.16** **	**1.36*** **
	[1.31]	**[−0.16]**	**[−1.51]**	**[0.41]**	**[0.32]**	**[−0.53]**	**[0.50]**	**[0.69]**	**[−3.27]**	**[0.68]**	**[2.20]**	**[1.75]**	**[2.07]**	**[2.57]**	**[3.04]**
β^{mkt}	−0.059	−0.25	−0.25	−0.25	−0.25	−0.25	−0.24	−0.25	−0.25	−0.25	−0.23	−0.26	−0.25	−0.17	−0.13
	[−0.15]	[−1.15]	[−1.13]	[−1.14]	[−1.13]	[−1.15]	[−1.11]	[−1.13]	[−1.14]	[−1.13]	[−1.06]	[−1.00]	[−0.97]	[−0.82]	[−0.66]
β^{hml}	−0.12	−0.080	−0.079	−0.080	−0.079	−0.079	−0.076	−0.079	−0.084	−0.081	−0.079	−0.045	−0.044	−0.091	−0.10
	[−0.83]	[−0.67]	[−0.66]	[−0.66]	[−0.66]	[−0.65]	[−0.63]	[−0.66]	[−0.70]	[−0.67]	[−0.66]	[−0.33]	[−0.32]	[−0.76]	[−0.84]
β^{smb}	0.046	0.033	0.034	0.034	0.033	0.032	0.036	0.035	0.028	0.033	0.052	0.061	0.067	0.066	0.079
	[0.44]	[0.44]	[0.45]	[0.46]	[0.44]	[0.43]	[0.49]	[0.47]	[0.38]	[0.45]	[0.74]	[0.77]	[0.85]	[0.91]	[1.09]
BM	0.19	0.046	0.046	0.046	0.046	0.045	0.036	0.045	0.049	0.049	0.034	0.065	0.062	0.043	0.043
	[1.27]	[1.08]	[1.10]	[1.09]	[1.08]	[1.06]	[0.84]	[1.06]	[1.18]	[1.13]	[0.82]	[1.29]	[1.21]	[1.02]	[1.03]
ln(Mcap)	−0.019	0.026	0.027	0.027	0.027	0.027	0.032	0.027	0.010	0.028	0.043	0.011	0.018	0.093	0.12*
	[−0.30]	[0.60]	[0.64]	[0.62]	[0.63]	[0.63]	[0.80]	[0.65]	[0.24]	[0.67]	[1.00]	[0.25]	[0.41]	[1.55]	[1.89]
DYD	0.16	−0.15	−0.17	−0.15	−0.17	−0.18	−0.18	−0.15	−0.17	−0.19	−0.18	−0.0020	0.0041	−0.23	−0.22
	[0.15]	[−0.28]	[−0.31]	[−0.29]	[−0.32]	[−0.34]	[−0.34]	[−0.28]	[−0.33]	[−0.35]	[−0.33]	[−0.00]	[0.01]	[−0.46]	[−0.44]
Id. Vol.	−0.19***	−0.21***	−0.21***	−0.21***	−0.21***	−0.21***	−0.21***	−0.21***	−0.19***	−0.20***	−0.21***	−0.25***	−0.25***	−0.19***	−0.18***
	[−2.82]	[−4.14]	[−4.14]	[−4.14]	[−4.14]	[−4.13]	[−4.23]	[−4.14]	[−3.93]	[−4.09]	[−4.21]	[−4.59]	[−4.59]	[−3.99]	[−3.84]
RET_{-1}	−0.69	−0.082	−0.084	−0.083	−0.068	−0.063	−0.070	−0.069	−0.11	−0.040	−0.080	−0.41	−0.44	−0.15	−0.21
	[−0.94]	[−0.16]	[−0.16]	[−0.16]	[−0.13]	[−0.12]	[−0.14]	[−0.13]	[−0.22]	[−0.08]	[−0.15]	[−0.72]	[−0.77]	[−0.29]	[−0.41]
$RET_{(-12,-2)}$	0.31*	0.17	0.16	0.17	0.17	0.17	0.17	0.17	0.16	0.16	0.19	0.19	0.21	0.21	0.23
	[1.87]	[1.04]	[1.01]	[1.03]	[1.04]	[1.04]	[1.06]	[1.03]	[1.01]	[1.02]	[1.26]	[1.08]	[1.18]	[1.29]	[1.40]
Observations	131,986†	360,626	360,626	360,626	360,626	360,626	360,066	360,624	360,626	360,624††	360,624††	294,284†††	294,284†††	360626	360626

Panel B: Loadings of ILMs in the cross-section of expected returns after orthogonalization relative to other liquidity measures

	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD	ILMT	ILMV
ILMT residual	0.18	1.22***	1.16**	1.17***	1.18**	1.18**	0.91*	1.16**	1.35***	1.06**	0.72	0.41	0.29	-	-
	[0.30]	[3.14]	[2.58]	[2.97]	[2.55]	[2.59]	[1.98]	[2.54]	[2.96]	[2.33]	[1.52]	[0.81]	[0.55]		
ILMV residual	0.26	1.45***	1.33***	1.40***	1.36***	1.38***	1.10**	1.34***	1.49***	1.25***	0.95**	0.59	0.48	-	-
	[0.42]	[3.79]	[3.03]	[3.60]	[3.00]	[3.09]	[2.43]	[2.97]	[3.32]	[2.82]	[2.05]	[1.16]	[0.92]		

† The number of observations reflects the largest sample of ANcerno data available from 2011-2014.

†† The number of observations reflects the largest sample available for ILLIQ and ILLIQ_OC.

††† The number of observations reflects the largest sample available for BBD and WBBD from 2010–2017.

\$1) definition of penny stocks, ILMs continue to explain the cross-section of expected returns. However, reflecting the relevance of alternative liquidity measures for smaller firms, the open-to-close version of Amihud's liquidity measure, ILLIQ_OC, also explains expected stock returns in the 2010-2019 period, consistent with Barardehi et al. (2021). In addition, the trade-time liquidity measures, BBD and $WBBD$, explain expected stock returns in the 2010–2017 period, consistent with Barardehi et al. (2019). However, realized institutional price impacts (InPrIM) no longer explain expected returns, a possible consequence of including stocks that institutional investors are reluctant or unable to hold.

In contrast, Table E.2 reports that with a stricter (under \$5) definition of penny stocks, which still excludes stocks held in limited amounts by institutional investors, ILMs and realized institutional price impacts explain the cross-section of returns. In addition, quoted depth has a negative coefficient, consistent with a characteristic liquidity premium, implying lower depth is associated with higher expected returns. In contrast, many standard liquidity measures, including spreads, Amihud, and trade-time measures, load with unexpected negative coefficients, indicating that such measures are unreliable liquidity measures for stocks more likely to be held by institutional investors. This reinforces the conclusion that standard liquidity measures are mostly relevant for small stocks.

Table E.2. Liquidity and the Cross-Section of Expected Stock Returns: 1-month ILMs. This table reports on the relation between alternative high-frequency liquidity measures and the cross-section of expected returns. In Panel A, equation (2) is estimated using liquidity measures ($LIQ_{j,m-2}$) constructed over 1-month horizons. Control variables include three-factor Fama-French betas ($\beta^{mkt}_{j,m-1}$, $\beta^{hml}_{j,m-1}$, $\beta^{smb}_{j,m-1}$), estimated using weekly observations from the two-year period ending in the final full week of month $m-1$, book-to-market ratio, ($\text{BM}_{j,m-1}$), natural log of market capitalization, ($\ln(\text{Mcap}_{j,m-1})$), dividend yield ($\text{DYD}_{j,m-1}$), defined as total dividends over the past 12 months divided by the share price at the end of month $m-1$, idiosyncratic volatility ($\text{IdVol}_{j,m-1}$), previous month's return ($RET_{(-1)}$), and preceding return from the prior 11 months ($RET_{(-12,-2)}$). Panel B replaces each high-frequency liquidity measure by the residuals of $ILMT$ and $ILMV$ with respect to each alternative liquidity measure, with residuals calculated separately for each monthly cross-section. Estimates are from Fama-MacBeth regressions that have Newey-West corrected standard errors with 6 lags. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$5. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Stock liquidity and the cross-section of expected returns

	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD	ILMT	ILMV
Constant	1.34	1.42	1.31	1.39	1.35	1.39	1.90**	1.37	1.70*	1.52*	1.66*	2.71***	2.64***	0.26	−0.46
	[1.22]	[1.64]	[1.55]	[1.59]	[1.61]	[1.64]	[2.18]	[1.61]	[1.98]	[1.76]	[1.84]	[3.01]	[2.93]	[0.23]	[−0.38]
Liquidity	**0.027****	**−0.068**	**−0.000011****	**−0.032**	**0.055**	**−0.070**	**−0.17****	**−0.024**	**−8.31*****	**−0.050**	**−0.25***	**−0.86*****	**−1.23*****	**0.67***	**0.88****
	[2.11]	**[−0.72]**	**[−2.06]**	**[−0.22]**	**[0.69]**	**[−0.68]**	**[−2.37]**	**[−0.33]**	**[−3.80]**	**[−0.91]**	**[−1.88]**	**[−3.62]**	**[−3.21]**	**[1.94]**	**[2.49]**
β^{mkt}	−0.0056	−0.11	−0.10	−0.11	−0.10	−0.10	−0.12	−0.11	−0.099	−0.11	−0.12	−0.13	−0.13	−0.055	−0.026
	[−0.01]	[−0.51]	[−0.49]	[−0.50]	[−0.48]	[−0.49]	[−0.56]	[−0.50]	[−0.46]	[−0.54]	[−0.58]	[−0.52]	[−0.50]	[−0.27]	[−0.13]
β^{hml}	−0.11	−0.11	−0.10	−0.11	−0.11	−0.11	−0.11	−0.11	−0.11	−0.11	−0.11	−0.057	−0.056	−0.11	−0.12
	[−0.74]	[−0.81]	[−0.78]	[−0.81]	[−0.81]	[−0.81]	[−0.80]	[−0.81]	[−0.87]	[−0.81]	[−0.82]	[−0.38]	[−0.37]	[−0.85]	[−0.92]
β^{smb}	0.12	0.036	0.035	0.037	0.038	0.036	0.023	0.038	0.039	0.026	0.016	0.00	0.0052	0.065	0.076
	[1.21]	[0.46]	[0.45]	[0.47]	[0.48]	[0.45]	[0.29]	[0.48]	[0.49]	[0.34]	[0.21]	[0.00]	[0.06]	[0.85]	[1.01]
BM	0.12	−0.0050	−0.0045	−0.0048	−0.0047	−0.0060	−0.012	−0.0053	−0.00030	0.000071	0.0013	0.054	0.050	−0.0071	−0.0045
	[0.94]	[−0.16]	[−0.14]	[−0.15]	[−0.15]	[−0.19]	[−0.37]	[−0.17]	[−0.01]	[0.00]	[0.04]	[1.09]	[1.02]	[−0.23]	[−0.14]
ln(Mcap)	0.0049	−0.0015	0.0040	−0.00	0.0015	0.00	−0.022	0.00075	−0.012	−0.0056	−0.012	−0.058	−0.054	0.043	0.069
	[0.11]	[−0.04]	[0.11]	[−0.01]	[0.04]	[0.00]	[−0.61]	[0.02]	[−0.34]	[−0.16]	[−0.31]	[−1.54]	[−1.45]	[0.97]	[1.43]
DYD	0.68	0.24	0.23	0.24	0.22	0.22	0.25	0.22	0.21	0.20	0.20	0.53	0.53	0.19	0.20
	[0.61]	[0.42]	[0.40]	[0.42]	[0.39]	[0.40]	[0.44]	[0.39]	[0.38]	[0.35]	[0.35]	[0.82]	[0.83]	[0.34]	[0.37]
Id. Vol.	−0.11	−0.18***	−0.18***	−0.18***	−0.18***	−0.18***	−0.17***	−0.18***	−0.17***	−0.18***	−0.17***	−0.14**	−0.14**	−0.17***	−0.17***
	[−1.52]	[−3.47]	[−3.48]	[−3.47]	[−3.48]	[−3.47]	[−3.21]	[−3.44]	[−3.34]	[−3.30]	[−3.18]	[−2.22]	[−2.26]	[−3.47]	[−3.44]
RET_{-1}	−0.80	−0.88	−0.87	−0.88	−0.87	−0.87	−0.86	−0.89	−0.89	−0.87	−0.85	−0.84	−0.85	−0.90	−0.92
	[−1.12]	[−1.49]	[−1.47]	[−1.49]	[−1.46]	[−1.46]	[−1.46]	[−1.49]	[−1.52]	[−1.47]	[−1.44]	[−1.24]	[−1.26]	[−1.50]	[−1.54]
$RET_{(-12,-2)}$	0.38*	0.17	0.17	0.17	0.17	0.17	0.15	0.17	0.18	0.17	0.16	0.13	0.13	0.21	0.23
	[1.89]	[1.10]	[1.10]	[1.09]	[1.09]	[1.11]	[1.00]	[1.10]	[1.16]	[1.07]	[1.02]	[0.68]	[0.68]	[1.34]	[1.45]
Observations	115,759†	297337	297337	297337	297337	297337	296805	297335	297337	297,335††	297,335††	242442	242442	297,337†††	297,337†††

Panel B: Loadings of ILMs in the cross-section of expected returns after orthogonalization relative to other liquidity measures

	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD	ILMT	ILMV
ILMT residual	−0.27	0.73**	0.64*	0.69**	0.64*	0.69**	0.88***	0.68*	0.84**	0.81**	0.93***	1.19***	1.14***	-	-
	[−0.54]	[2.55]	[1.90]	[2.46]	[1.93]	[2.04]	[2.70]	[1.92]	[2.46]	[2.50]	[2.90]	[3.02]	[2.97]		
ILMV residual	−0.22	0.96***	0.84**	0.92***	0.85**	0.90**	1.03***	0.88**	1.00***	0.97***	1.06***	1.23***	1.18***	-	-
	[−0.47]	[3.28]	[2.41]	[3.20]	[2.51]	[2.62]	[3.14]	[2.44]	[2.82]	[3.04]	[3.45]	[3.23]	[3.20]		

† The number of observations reflects the largest sample of ANcerno data available from 2011-2014.

†† The number of observations reflects the largest sample available for ILLIQ and ILLIQ_OC.

††† The number of observations reflects the largest sample available for BBD and WBBD from 2010–2017.

Panel B in Tables E.1 and E.2 highlights the incremental information content of $ILMT$ and $ILMV$ with respect to each alternative liquidity measure. First, the residuals of each ILM with respected an alternative measure are calculated using Fama-MacBeth regressions. These residuals are then used as LIQ in equation (2). For both minimum price filters, with the exception of realized institutional price impacts (InPrIM), ILM residuals explain the cross-section of two-months-ahead returns whenever the liquidity measure against which these residuals are calculated does not explain the cross-section of these returns (with expected sign) in Panel A. As such, our findings provide unambiguous evidence that ILMs outperform all existing liquidity measures in explaining the cross-section of expected returns.[79]

F Portfolio Sorts: Alternative Liquidity Measures

This section employs simple portfolio sorts to compare the economic magnitudes of the premia associated with all liquidity measures used in our study. We sort each monthly cross-section into ten portfolios (deciles) of each liquidity measure (LIQ). We then calculate average monthly stock returns of each portfolio as well as monthly returns associated with four long-short strategies that buy illiquid stocks and sell liquid stocks. Strategy (1) in long on decile 7 and short on decile 4; strategy (2) is long on decile 8 and short on decile 3; strategy (3) is long on decile 9 and short on decile 2; and the "traditional" strategy (4) is long on decile 10 and short on decile (1). Examining these four strategies reveals whether liquidity premia are only attributable to the tails of the distributions. We obtain three-factor alphas by regressing the time series of portfolio returns as well as those of the long-short strategies on Fama-French three factors. We conduct three versions of these analyses based on samples with minimum previous month's end share price filters of \$1, \$2, and \$5.[80]

Table F.1 reports that ILMs are the only measures for which the traditional long-short strategy (4) consistently produces three-factor liquidity premia of nearly 1% or higher. In addition, $ILMV$ is the sole liquidity measure for which all four long-short strategies produce significant liquidity premia. This finding indicates that $ILMV$ identifies economically relevant differences in stock liquidity even for stocks with intermediate trading costs, highlighting the practical relevance of

[79] In untabulated results, we verify that the converse is not true.

[80] Note that the findings regarding $ILMT$ and $ILMV$ match those reported in Panels A–C in Table 10.

ILMs. Long-short strategies based on dollar quoted, effective, and realized spreads also produce relatively consistent liquidity premia. However, these measures are impacted by variations in share price: ceteris paribus, higher share price is associated with wider spreads measures. This observation is consistent with the finding that long-short strategies based on percentage (relative) quoted, effective, and realized spreads do *not* produce significant three-factor alphas. That is, when adjusted for share price, these spreads-based measures fail to capture liquidity. This interpretation is reinforced by the regression analyses reported in Tables 8, E.1, and E.2 where controlling for other stock characteristics, including book-to-market ratio and market-capitalization, renders all spread-based measures insignificant predictors of expected returns.

Table F.1. **Liquidity Alphas:** This table presents three-factor alphas of liquidity measures ($LIQ_{j,m-2}$) from 1-month horizons. Every month, stocks are sorted into deciles of the respective LIQ. Alphas for four long-short strategies are reported: long decile 7, short decile 4; long decile 8, short decile 3; long decile 9, short decile 2; and long decile 10, short decile 1. The 118-month time-series of monthly average portfolio returns for each portfolio (net of 1-month T-bill rate) and the long-short strategies are regressed on the Fama-French three factors to obtain alphas. The sample period is from 2010–2019, excluding stocks with previous month-end's closing price below \$1, \$2, and \$5, in Panels A, B, and C, respectively. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: \$1 minimum share price

	Liquidity portfolios								Long-short strategies			
LIQ	1	2	3	4	7	8	9	10	7−4	8−3	9−2	10−1
InPrIm	−0.14	0.082	0.058	−0.042	0.064	0.17	0.072	0.014	0.11	0.11	−0.0098	0.15
	[−0.58]	[0.63]	[0.48]	[−0.23]	[0.54]	[1.40]	[0.64]	[0.07]	[0.47]	[0.53]	[−0.06]	[0.91]
QSP	−0.45***	−0.48***	−0.24*	−0.16*	0.10	0.13	0.37***	0.40***	0.26**	0.37***	0.85***	0.85***
	[−3.17]	[−3.73]	[−1.94]	[−1.80]	[1.24]	[1.44]	[3.87]	[3.44]	[2.00]	[3.72]	[5.39]	[4.00]
ShrDepth†	−0.15* −	0.21***	−0.13	−0.21***	0.041	0.28*	0.32*	0.78***	0.25*	0.41**	0.53**	0.93***
	[1.79]	[2.83]	[1.59]	[3.26]	[−0.34]	[−1.89]	[−1.78]	[−4.07]	[−1.76]	[−2.11]	[−2.52]	[−4.03]
EFSP	−0.57***	−0.28***	−0.39***	−0.23***	0.13	0.16*	0.27**	0.47***	0.35***	0.56***	0.56***	1.05***
	[−3.29]	[−2.66]	[−4.03]	[−3.76]	[1.36]	[1.74]	[2.59]	[4.40]	[3.59]	[5.47]	[4.19]	[4.54]
RESP	−0.14	−0.28***	−0.23***	−0.31***	−0.082	0.11	0.30***	0.37***	0.23*	0.34***	0.58***	0.51***
	[−1.03]	[−2.90]	[−3.07]	[−2.99]	[−0.86]	[1.18]	[2.68]	[3.14]	[1.94]	[3.14]	[4.02]	[2.80]
PIMP	−0.62***	−0.33***	−0.32***	−0.27***	0.16**	0.17**	0.33***	0.32***	0.43***	0.49***	0.66***	0.94***
	[−3.21]	[−2.66]	[−3.26]	[−3.65]	[2.39]	[2.50]	[3.65]	[3.34]	[4.91]	[4.50]	[4.31]	[4.87]
Lambda	0.14**	−0.016	−0.12*	0.075	0.021	0.046	−0.32*	−0.34	−0.054	0.17	−0.30	−0.49
	[2.61]	[−0.18]	[−1.79]	[1.15]	[0.25]	[0.44]	[−1.78]	[−1.15]	[−0.58]	[1.20]	[−1.52]	[−1.60]
AMVST	−0.36***	−0.20***	−0.11**	−0.17***	0.013	−0.13	0.29*	0.41**	0.19**	−0.015	0.49***	0.77***
	[−3.16]	[−2.83]	[−2.17]	[−2.99]	[0.17]	[−1.10]	[1.91]	[2.11]	[2.25]	[−0.13]	[3.06]	[3.76]
ROLL	−0.16*	−0.12	−0.18**	0.085	0.22***	0.082	−0.20	−0.69***	0.14	0.26**	−0.075	−0.53**
	[−1.70]	[−1.35]	[−2.44]	[1.09]	[3.64]	[0.76]	[−1.57]	[−2.83]	[1.28]	[2.28]	[−0.55]	[−2.45]
ILLIQ	0.040	−0.081	−0.11	0.031	−0.11	−0.26**	−0.16	0.32	−0.14	−0.15	−0.078	0.28
	[0.82]	[−0.88]	[−1.34]	[0.52]	[−1.28]	[−2.24]	[−0.85]	[1.17]	[−1.43]	[−1.02]	[−0.38]	[1.03]
ILLIQ_OC	0.048	−0.099	−0.089	−0.00036	−0.100	−0.25**	−0.065	0.21	−0.099	−0.16	0.034	0.16
	[0.94]	[−1.09]	[−1.03]	[−0.01]	[−1.08]	[−2.31]	[−0.36]	[0.75]	[−0.92]	[−1.12]	[0.16]	[0.57]
BBD	0.049	0.026	−0.13	0.067	0.021	−0.063	−0.013	−0.011	−0.046	0.065	−0.038	−0.059
	[1.14]	[0.25]	[−1.59]	[1.38]	[0.21]	[−0.51]	[−0.08]	[−0.03]	[−0.41]	[0.39]	[−0.20]	[−0.18]
WBBD	0.036	0.030	−0.13*	0.097*	0.015	0.0040	−0.048	0.0014	−0.081	0.14	−0.078	−0.035
	[0.80]	[0.29]	[−1.70]	[1.86]	[0.16]	[0.03]	[−0.28]	[0.00]	[−0.73]	[0.80]	[−0.40]	[−0.11]
ILMT	−0.32***	−0.34***	−0.19**	−0.17	−0.032	0.089	0.38**	0.64***	0.14	0.28	0.72***	0.96***
	[−2.77]	[−3.82]	[−2.13]	[−1.58]	[−0.30]	[0.63]	[2.48]	[4.25]	[0.86]	[1.62]	[3.72]	[4.30]
ILMV	−0.63***	−0.44***	−0.25***	−0.25***	−0.027	0.32***	0.32**	0.64***	0.22**	0.57***	0.77***	1.27***
	[−4.28]	[−4.40]	[−2.88]	[−3.56]	[−0.28]	[2.85]	[2.10]	[4.76]	[2.15]	[4.17]	[4.28]	[5.49]

Continued on next page

Table F.1 – *continued from previous page*

Panel B: $2 minimum share price

	Liquidity portfolios								Long-short strategies			
LIQ	1	2	3	4	7	8	9	10	7−4	8−3	9−2	10−1
InPrIm	−0.092	0.066	0.12	−0.055	0.053	0.13	0.078	0.23	0.11	0.0077	0.013	0.32**
	[−0.42]	[0.51]	[1.22]	[−0.32]	[0.44]	[1.13]	[0.65]	[1.10]	[0.51]	[0.05]	[0.08]	[2.31]
QSP	−0.41***	−0.26**	−0.21**	−0.21***	0.098	0.14	0.34***	0.41***	0.30**	0.35***	0.60***	0.82***
	[−3.41]	[−2.47]	[−1.99]	[−2.63]	[1.15]	[1.64]	[3.48]	[3.83]	[2.54]	[3.51]	[3.71]	[4.28]
ShrDepth†	−0.15*	−0.19***	−0.14*	−0.22***	0.0090	0.24*	0.29**	0.56***	0.23	0.38**	0.48***	0.71***
	[1.72]	[2.72]	[1.68]	[3.00]	[−0.07]	[−1.74]	[−2.25]	[−4.19]	[−1.52]	[−2.17]	[−2.90]	[−3.92]
EFSP	−0.47***	−0.21**	−0.33***	−0.11*	0.061	0.21**	0.29***	0.42***	0.17	0.54***	0.51***	0.89***
	[−3.16]	[−2.06]	[−4.44]	[−1.70]	[0.70]	[2.33]	[2.99]	[3.87]	[1.53]	[5.71]	[3.52]	[4.08]
RESP	−0.18	−0.23**	−0.23***	−0.19**	−0.075	0.097	0.33***	0.42***	0.12	0.33***	0.56***	0.60***
	[−1.51]	[−2.57]	[−3.12]	[−2.59]	[−0.98]	[1.09]	[3.11]	[3.54]	[1.24]	[2.91]	[4.07]	[3.15]
PIMP	−0.42***	−0.28**	−0.24***	−0.13*	0.15**	0.24***	0.29***	0.26***	0.28***	0.48***	0.57***	0.68***
	[−2.68]	[−2.57]	[−2.68]	[−1.72]	[2.48]	[3.20]	[3.15]	[2.81]	[2.84]	[4.44]	[3.85]	[3.63]
Lambda	0.13**	−0.016	−0.14*	0.027	0.090	0.17*	−0.20	−0.28	0.063	0.31**	−0.18	−0.41
	[2.42]	[−0.20]	[−1.92]	[0.36]	[1.17]	[1.81]	[−1.55]	[−1.10]	[0.67]	[2.17]	[−1.11]	[−1.54]
AMVST	−0.37***	−0.20**	−0.048	−0.18***	0.058	0.0034	0.22**	0.43**	0.24**	0.052	0.42***	0.80***
	[−3.12]	[−2.57]	[−1.05]	[−3.33]	[0.63]	[0.04]	[2.10]	[2.45]	[2.34]	[0.55]	[3.13]	[4.22]
ROLL	−0.12	−0.12	−0.19**	0.099	0.31***	0.14*	−0.055	−0.76***	0.21*	0.33***	0.063	−0.64***
	[−1.34]	[−1.54]	[−2.58]	[1.13]	[4.36]	[1.90]	[−0.50]	[−3.91]	[1.70]	[3.71]	[0.59]	[−3.20]
ILLIQ	0.040	−0.058	−0.15*	0.030	−0.013	−0.073	−0.050	0.20	−0.043	0.076	0.0081	0.16
	[0.81]	[−0.67]	[−1.85]	[0.49]	[−0.17]	[−0.62]	[−0.31]	[0.88]	[−0.53]	[0.47]	[0.04]	[0.69]
ILLIQ_OC	0.041	−0.071	−0.095	−0.036	0.0036	−0.10	0.023	0.14	0.040	−0.0085	0.094	0.10
	[0.83]	[−0.76]	[−1.19]	[−0.62]	[0.04]	[−0.93]	[0.16]	[0.61]	[0.42]	[−0.06]	[0.51]	[0.43]
BBD	0.040	0.057	−0.15*	0.10	−0.072	0.13	0.051	−0.062	−0.18	0.28	−0.0052	−0.10
	[0.91]	[0.55]	[−1.77]	[1.56]	[−0.83]	[0.91]	[0.44]	[−0.23]	[−1.41]	[1.45]	[−0.03]	[−0.38]
WBBD	0.047	0.053	−0.16*	0.090	−0.052	0.16	0.093	−0.11	−0.14	0.31	0.040	−0.16
	[1.07]	[0.52]	[−1.78]	[1.40]	[−0.59]	[1.10]	[0.82]	[−0.39]	[−1.19]	[1.64]	[0.22]	[−0.55]
ILMT	−0.30***	−0.33***	−0.21**	−0.062	0.023	0.11	0.34**	0.62***	0.085	0.31*	0.67***	0.93***
	[−2.70]	[−4.05]	[−2.17]	[−0.82]	[0.27]	[0.92]	[2.54]	[4.48]	[0.72]	[1.81]	[4.32]	[4.33]
ILMV	−0.58***	−0.33***	−0.23***	−0.25***	0.041	0.28***	0.31**	0.63***	0.30***	0.50***	0.65***	1.20***
	[−3.97]	[−3.86]	[−2.76]	[−3.68]	[0.59]	[3.37]	[2.26]	[4.97]	[3.10]	[4.27]	[3.72]	[5.09]

Continued on next page

Table F.1 – *continued from previous page*

Panel C: \$5 minimum share price

	Liquidity portfolios								Long-short strategies			
LIQ	1	2	3	4	7	8	9	10	7−4	8−3	9−2	10−1
InPrIm	0.080	0.21*	−0.017	−0.060	0.041	0.17	0.11	0.28**	0.10	0.19	−0.095	0.20
	[0.40]	[1.77]	[−0.14]	[−0.33]	[0.34]	[1.37]	[1.01]	[2.09]	[0.50]	[1.00]	[−0.58]	[1.35]
QSP	−0.23***	−0.13	−0.056	−0.019	0.071	0.21**	0.39***	0.41***	0.090	0.27**	0.52***	0.65***
	[−2.73]	[−1.58]	[−0.61]	[−0.31]	[0.82]	[2.55]	[4.13]	[3.92]	[0.86]	[2.36]	[3.49]	[3.98]
ShrDepth†	−0.13	−0.23***	−0.18**	−0.13**	−0.20***	−0.036	0.11	0.18**	0.069	0.14	0.34**	0.31*
	[1.31]	[3.04]	[2.03]	[2.00]	[3.06]	[0.32]	[−1.06]	[−1.99]	[0.72]	[−0.99]	[−2.39]	[−1.88]
EFSP	−0.24**	−0.11	−0.15**	0.026	0.15*	0.22***	0.31***	0.48***	0.13	0.37***	0.41***	0.72***
	[−2.12]	[−1.30]	[−2.58]	[0.44]	[1.81]	[2.74]	[3.27]	[4.36]	[1.26]	[3.66]	[2.93]	[3.79]
RESP	−0.10	−0.063	−0.17**	−0.080	0.047	0.21**	0.39***	0.52***	0.13	0.38***	0.46***	0.62***
	[−0.95]	[−0.96]	[−2.57]	[−1.25]	[0.69]	[2.41]	[3.53]	[4.38]	[1.38]	[3.26]	[3.17]	[3.12]
PIMP	−0.079	−0.19**	−0.044	−0.039	0.15**	0.20***	0.31***	0.33***	0.19*	0.25**	0.50***	0.41**
	[−0.84]	[−2.03]	[−0.67]	[−0.50]	[2.31]	[2.66]	[3.69]	[3.16]	[1.81]	[2.52]	[3.87]	[2.56]
Lambda	0.14***	0.0072	−0.15*	−0.025	0.15**	0.13	0.32***	0.011	0.18*	0.28**	0.31**	−0.13
	[2.71]	[0.09]	[−1.67]	[−0.33]	[2.43]	[1.60]	[3.03]	[0.06]	[1.85]	[2.04]	[2.00]	[−0.66]
AMVST	−0.30**	−0.13*	0.043	−0.036	0.057	0.28***	0.30***	0.55***	0.093	0.24**	0.43***	0.85***
	[−2.32]	[−1.84]	[0.73]	[−0.65]	[0.86]	[3.79]	[2.75]	[4.69]	[1.11]	[2.48]	[3.26]	[4.11]
ROLL	−0.058	0.072	0.00013	0.13**	0.26***	0.27***	0.049	−0.46***	0.13	0.27***	−0.023	−0.40***
	[−0.82]	[1.10]	[0.00]	[2.12]	[4.20]	[5.24]	[0.55]	[−3.61]	[1.41]	[2.73]	[−0.21]	[−3.23]
ILLIQ	0.045	−0.039	−0.11	−0.048	0.085	0.12	0.26**	0.44***	0.13	0.23*	0.30*	0.39**
	[0.92]	[−0.43]	[−1.48]	[−0.71]	[1.13]	[1.31]	[2.08]	[2.73]	[1.23]	[1.69]	[1.67]	[2.13]
ILLIQ_OC	0.045	−0.036	−0.093	−0.059	0.11	0.12	0.25**	0.45***	0.16	0.21	0.28	0.40**
	[0.90]	[−0.48]	[−1.04]	[−0.88]	[1.28]	[1.55]	[2.01]	[2.74]	[1.43]	[1.62]	[1.65]	[2.16]
BBD	0.071*	0.045	−0.12	−0.030	0.12	0.11	0.31**	0.39**	0.15	0.23	0.26	0.32*
	[1.67]	[0.51]	[−1.20]	[−0.40]	[1.66]	[1.20]	[2.21]	[2.55]	[1.27]	[1.38]	[1.34]	[1.96]
WBBD	0.062	0.050	−0.14	−0.015	0.13*	0.16	0.27*	0.42***	0.14	0.30*	0.22	0.36**
	[1.44]	[0.56]	[−1.38]	[−0.21]	[1.74]	[1.53]	[1.91]	[2.80]	[1.26]	[1.67]	[1.11]	[2.23]
ILMT	−0.29***	−0.24***	−0.14*	0.053	0.12	0.28***	0.38***	0.65***	0.067	0.42***	0.62***	0.95***
	[−2.66]	[−2.89]	[−1.98]	[0.78]	[1.25]	[2.84]	[3.49]	[4.73]	[0.56]	[3.25]	[4.39]	[4.30]
ILMV	−0.43***	−0.21***	−0.14**	−0.11	0.19***	0.37***	0.43***	0.68***	0.30***	0.51***	0.64***	1.10***
	[−3.35]	[−2.64]	[−2.16]	[−1.54]	[2.86]	[4.65]	[4.02]	[5.32]	[3.64]	[4.44]	[3.92]	[4.82]

† For consistency, returns to long-short strategies based on quoted depth (ShrDepth) are multiplied by −1.

G Portfolio double-sorts

This section provides return differences between stocks falling in different levels of ILM and stock characteristics. Double sorts based on ILMs and other stock characteristics provide additional evidence that the 3-factor risk-adjusted portfolio return spreads associated with our liquidity measures are not concentrated in specific subsets of stocks. These double sorts control for market beta, market capitalization, book-to-market ratios, past returns, and the share of sub-penny volume. After excluding stocks priced below \$5 at the end of the preceding month, we form an array of 5×5 portfolios that first condition on a stock characteristic, and then on an ILM.[81] Next, we estimate monthly portfolio returns as well as return spreads between the most and least liquid stock portfolios, conditional on the level of each stock characteristic.

Table G.1 documents liquidity premia for high- and low-beta, small and large, growth and value stocks, past losers and past winners, and stocks with low and high sub-penny executed volume. A slightly smaller liquidity premia is apparent among large stocks, past winners, and value stocks. However, reflecting lowered measurement error, the significant liquidity premia grows by nearly six times as the share of sub-penny executed volume rises from its bottom to its top quintile. Online Appendix H establishes the robustness of these findings to constructing ILMs over 3-month rolling windows. Therefore, the liquidity premia associated with ILMs are largely orthogonal to stock characteristics known to influence expected returns.

Finally, we investigate whether trading costs can explain the returns of anomalies based on stock characteristics by changing the order of the double sorts—first conditioning on a ILM, and then on a stock characteristic. Table G.2 reports evidence that low-beta and value premia are present in both liquid and illiquid stocks. In contrast, momentum's alpha is only significant among the 20% least liquid stocks, suggesting that momentum profits do not survive institutional trading costs (Lesmond et al. (2004); Korajczyk and Sadka (2004)).[82]

[81] Our choice of the \$5 minimum share price precludes effects attributable to penny stocks, leading to conservative estimates. Qualitative findings are unaffected by using \$1 and \$2 share price filters.

[82] Online Appendix H confirms results are robust to constructing ILMs over 3-month rolling windows.

Table G.1. Portfolio Alphas: Stock Characteristic and ILM Double-Sorts. This table presents three-factor alphas using CRSP breakpoints. Stocks are first sorted into stock characteristic quintiles $X \in \{\beta^{mkt}, \text{Mcap}, RET_{(-12,-2)}, \text{BM}, \text{SPVS}\}$. Within each characteristic quintile, stocks are further sorted into $LIQ \in \{ILMT, ILMV\}$ quintiles. Monthly 5×5 portfolio returns are equally-weighted averages of monthly stock returns in the portfolio. The time-series returns of each portfolio (after subtracting the 1-month Treasury-bill rate) including the long-short portfolio are then regressed on Fama-French three factors. The resulting intercepts are three-factor alphas. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below $5. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Sequential double sorts on market beta and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of market beta	Low	0.23	−0.011	0.41**	0.75***	0.82***	0.59***	−0.069	0.19	0.50***	0.74***	0.82***	0.89***
		[1.47]	[−0.09]	[2.58]	[5.07]	[4.90]	[2.74]	[−0.41]	[1.52]	[4.20]	[4.58]	[5.01]	[3.94]
	2	0.021	0.32**	0.57***	0.47***	0.47***	0.44***	0.13	0.32***	0.45***	0.47***	0.49***	0.37**
		[0.20]	[2.61]	[6.30]	[4.91]	[3.58]	[2.91]	[1.11]	[3.15]	[5.05]	[4.40]	[3.74]	[2.12]
	3	0.059	−0.066	0.073	0.30***	0.30**	0.24	−0.12	0.038	0.079	0.27**	0.39***	0.50***
		[1.08]	[−0.72]	[0.70]	[2.80]	[2.40]	[1.60]	[−1.62]	[0.47]	[0.84]	[2.61]	[3.79]	[3.90]
	4	−0.19*	−0.15	−0.011	−0.13	0.14	0.33**	−0.34***	−0.10	−0.19*	0.12	0.18	0.52***
		[−1.90]	[−1.50]	[−0.10]	[−1.02]	[0.84]	[1.99]	[−3.94]	[−1.07]	[−1.69]	[1.07]	[1.08]	[3.56]
	High	−0.78***	−0.54**	−0.39**	−0.38**	−0.22	0.57**	−0.86***	−0.39**	−0.59***	−0.31**	−0.16	0.70**
		[−2.99]	[−2.55]	[−2.39]	[−2.23]	[−1.34]	[2.03]	[−2.86]	[−2.21]	[−2.81]	[−2.31]	[−1.03]	[2.51]

Panel B: Sequential double sorts on market capitalization and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of market capitalization	Low	−0.69***	−0.0053	0.42***	0.70***	0.76***	1.45***	−0.87***	0.20	0.37**	0.68***	0.79***	1.67***
		[−2.96]	[−0.03]	[2.82]	[4.08]	[4.45]	[5.23]	[−3.90]	[1.07]	[2.32]	[4.23]	[4.61]	[6.06]
	2	−0.76***	−0.093	0.33***	0.50***	0.46***	1.22***	−0.90***	−0.025	0.31***	0.54***	0.51***	1.41***
		[−4.73]	[−0.66]	[3.16]	[3.94]	[2.72]	[4.92]	[−4.85]	[−0.18]	[3.08]	[3.73]	[3.18]	[5.29]
	3	−0.35***	0.14	0.091	0.25***	0.28**	0.63***	−0.33**	−0.079	0.24**	0.23**	0.35***	0.68***
		[−3.56]	[1.41]	[0.85]	[2.65]	[2.48]	[3.90]	[−2.49]	[−0.91]	[2.37]	[2.14]	[3.15]	[3.32]
	4	−0.35*	−0.14	0.14	0.052	0.10	0.45**	−0.52**	−0.055	0.054	0.059	0.27***	0.79***
		[−1.92]	[−1.05]	[1.47]	[0.55]	[1.45]	[2.36]	[−2.53]	[−0.45]	[0.65]	[0.62]	[3.62]	[3.82]
	High	−0.28***	0.024	0.10*	0.13	0.23***	0.50***	−0.25**	0.075	0.11	0.052	0.22***	0.47***
		[−2.86]	[0.34]	[1.71]	[1.51]	[3.92]	[4.78]	[−1.98]	[1.29]	[1.45]	[0.50]	[2.71]	[3.52]

Continued on next page

Table G.1 – *continued from previous page*

Panel C: Sequential double sorts on book-to-market ratio and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of book-to-market ratio	Low	−0.13	−0.14	0.065	0.012	0.26	0.38	−0.32*	−0.029	−0.063	0.14	0.34*	0.65***
		[−0.98]	[−0.98]	[0.40]	[0.08]	[1.19]	[1.52]	[−1.92]	[−0.25]	[−0.53]	[0.83]	[1.78]	[3.27]
	2	−0.29**	−0.15	0.12	−0.080	0.13	0.42*	−0.37***	−0.016	−0.16	0.075	0.19	0.56***
		[−2.10]	[−1.39]	[0.96]	[−0.63]	[0.94]	[1.95]	[−2.65]	[−0.14]	[−1.36]	[0.65]	[1.58]	[2.89]
	3	−0.22**	−0.057	−0.043	0.11	0.088	0.31*	−0.31**	−0.13	0.013	0.15	0.15	0.46**
		[−2.22]	[−0.49]	[−0.55]	[0.94]	[0.62]	[1.68]	[−2.60]	[−1.20]	[0.17]	[1.12]	[1.15]	[2.41]
	4	−0.36***	0.053	0.15	0.34**	0.66***	1.02***	−0.43***	−0.017	0.18**	0.46***	0.65***	1.08***
		[−3.22]	[0.45]	[1.35]	[2.47]	[4.27]	[4.48]	[−3.36]	[−0.13]	[2.08]	[3.09]	[4.21]	[4.63]
	High	−0.32*	0.020	0.26	0.69***	0.88***	1.20***	−0.43**	0.11	0.24	0.75***	0.87***	1.29***
		[−1.90]	[0.13]	[1.45]	[4.41]	[5.35]	[4.15]	[−2.04]	[0.76]	[1.61]	[5.38]	[5.33]	[4.18]

Panel D: Sequential double sorts on past 11-month return and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of past return	Low	−0.93***	−0.56***	−0.27	−0.18	−0.038	0.89***	−1.00***	−0.61***	−0.26	−0.025	−0.075	0.93**
		[−3.55]	[−2.82]	[−1.25]	[−0.95]	[−0.21]	[2.70]	[−3.22]	[−3.14]	[−1.60]	[−0.15]	[−0.40]	[2.37]
	2	−0.056	−0.12	0.14	0.25*	0.57***	0.63***	−0.17	0.036	0.11	0.23*	0.57***	0.74***
		[−0.44]	[−0.96]	[1.05]	[1.96]	[4.26]	[3.22]	[−1.46]	[0.33]	[0.86]	[1.87]	[4.16]	[3.83]
	3	−0.081	0.22**	0.30***	0.34***	0.93***	1.01***	−0.085	0.16*	0.15	0.53***	0.94***	1.02***
		[−1.16]	[2.24]	[2.77]	[2.67]	[6.61]	[5.81]	[−1.08]	[1.76]	[1.39]	[4.18]	[6.64]	[6.16]
	4	−0.022	0.15	0.088	0.35***	0.74***	0.76***	0.013	0.042	0.14	0.44***	0.68***	0.67***
		[−0.24]	[1.51]	[0.78]	[3.14]	[5.23]	[4.54]	[0.13]	[0.34]	[1.42]	[4.31]	[4.59]	[3.45]
	High	−0.21	−0.21	0.0078	0.23	0.40**	0.61***	−0.40*	−0.10	−0.18	0.27*	0.63***	1.03***
		[−1.03]	[−1.06]	[0.05]	[1.64]	[2.44]	[2.90]	[−1.92]	[−0.53]	[−1.08]	[1.86]	[3.84]	[4.21]

Continued on next page

Table G.1 – *continued from previous page*

Panel E: Sequential double sorts on share of sub-penny trade volume and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of sub-penny volume	Low	0.033	0.037	0.20**	0.17*	0.38***	0.35*	0.058	0.029	0.18**	0.14*	0.42***	0.36**
		[0.32]	[0.43]	[2.39]	[1.73]	[3.19]	[1.98]	[0.56]	[0.33]	[2.36]	[1.71]	[3.62]	[2.01]
	2	0.051	0.10	0.11	0.17***	0.38***	0.33*	−0.013	0.18*	0.087	0.15**	0.41***	0.42***
		[0.59]	[0.96]	[1.18]	[2.65]	[3.46]	[1.94]	[−0.17]	[1.88]	[1.00]	[2.05]	[3.25]	[2.65]
	3	−0.11	−0.084	−0.070	0.10	0.46***	0.57***	−0.12	−0.11	−0.11	0.15	0.48***	0.60***
		[−1.17]	[−0.87]	[−0.73]	[0.81]	[3.70]	[3.44]	[−1.11]	[−1.12]	[−1.15]	[1.52]	[3.92]	[3.25]
	4	−0.12	−0.15	−0.010	0.27**	0.58***	0.70***	−0.15	−0.10	−0.0014	0.23*	0.59***	0.75***
		[−1.27]	[−1.11]	[−0.07]	[2.11]	[3.14]	[2.94]	[−1.27]	[−0.84]	[−0.01]	[1.67]	[3.81]	[3.26]
	High	−1.17***	−0.64***	−0.053	0.56***	0.82***	1.99***	−1.15***	−0.81***	0.093	0.57***	0.83***	1.98***
		[−5.07]	[−3.55]	[−0.32]	[2.93]	[4.87]	[6.20]	[−4.94]	[−4.91]	[0.49]	[2.75]	[4.88]	[6.01]

Table G.2. Portfolio Alphas: ILM and Stock Characteristic Double-Sorts. This table presents three-factor alphas using CRSP breakpoints. Stocks are sorted into liquidity quintiles based on $LIQ \in \{ILMT, ILMV\}$. Within each liquidity quintile, stocks are further sorted into stock characteristic quintiles $X \in \{\beta^{mkt}, \text{Mcap}, RET_{(-12,-2)}, \text{BM}, \}$. Monthly 5×5 portfolio returns are equally-weighted averages of monthly stock returns in the portfolio. The time-series returns of each portfolio (after subtracting the 1-month Treasury-bill rate) including the long-short portfolio are then regressed on Fama-French three factors. The resulting intercepts are three-factor alphas. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$5. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Sequential double sorts on $ILMT$ and stock characteristics

		Portfolios of beta						Portfolios of market capitalization					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of $ILMT$	Low	0.048	0.031	−0.11	−0.41***	−0.87***	−0.92**	−0.85***	−0.37**	−0.053	−0.021	−0.030	0.82***
		[0.44]	[0.37]	[−1.27]	[−2.69]	[−3.01]	[−2.57]	[−3.99]	[−2.33]	[−0.43]	[−0.21]	[−0.72]	[3.77]
	2	0.32*	0.18**	0.034	−0.18*	−0.57***	−0.89***	−0.33**	−0.14	0.029	0.012	0.20***	0.54***
		[1.76]	[2.15]	[0.35]	[−1.73]	[−2.79]	[−2.66]	[−2.24]	[−1.17]	[0.24]	[0.11]	[2.95]	[3.05]
	3	0.14	0.26***	0.12	−0.051	−0.43**	−0.57**	−0.34**	0.029	0.15	0.12	0.065	0.40**
		[1.34]	[2.68]	[1.07]	[−0.50]	[−2.17]	[−2.25]	[−2.09]	[0.27]	[1.42]	[1.53]	[0.82]	[2.03]
	4	0.26**	0.54***	0.36***	0.016	−0.18	−0.44**	−0.30	0.47***	0.30***	0.37***	0.16**	0.46*
		[2.07]	[5.28]	[3.47]	[0.12]	[−1.05]	[−1.99]	[−1.29]	[4.06]	[3.39]	[3.69]	[2.00]	[1.74]
	High	0.71***	0.81***	0.47***	0.44***	0.16	−0.56**	0.29	0.80***	0.59***	0.45***	0.46***	0.18
		[3.49]	[5.99]	[3.24]	[3.54]	[1.09]	[−2.21]	[1.41]	[4.23]	[4.11]	[2.74]	[3.44]	[0.71]

		Portfolios of book-to-market ratio						Portfolios of past return ($R_{(-12,-2)}$)					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of $ILMT$	Low	−0.11	−0.23**	−0.32**	−0.27**	−0.39***	−0.28	−0.84***	−0.017	−0.075	−0.090	−0.30	0.54
		[−0.67]	[−2.06]	[−2.59]	[−2.52]	[−3.15]	[−1.42]	[−3.33]	[−0.14]	[−1.05]	[−0.83]	[−1.54]	[1.56]
	2	0.12	0.036	−0.019	−0.23*	−0.13	−0.26	−0.60***	0.078	0.24**	0.22**	−0.17	0.43
		[0.68]	[0.41]	[−0.20]	[−1.95]	[−0.81]	[−0.94]	[−2.96]	[0.68]	[2.61]	[2.25]	[−0.79]	[1.27]
	3	−0.059	−0.067	0.041	−0.019	0.13	0.19	−0.35	0.083	0.19*	0.11	−0.012	0.34
		[−0.41]	[−0.60]	[0.37]	[−0.20]	[0.87]	[0.82]	[−1.65]	[0.63]	[1.84]	[0.91]	[−0.08]	[1.09]
	4	0.16	0.18**	0.12	0.31***	0.22	0.068	−0.24	0.14	0.37***	0.43***	0.29**	0.52*
		[1.04]	[2.09]	[1.06]	[2.94]	[1.21]	[0.35]	[−0.94]	[1.20]	[2.81]	[3.88]	[2.06]	[1.72]
	High	0.18	0.18	0.65***	0.84***	0.74***	0.56**	−0.15	0.51***	0.90***	0.74***	0.59***	0.74***
		[0.99]	[1.29]	[4.18]	[5.10]	[3.92]	[2.07]	[−0.80]	[3.66]	[6.97]	[4.96]	[4.49]	[3.54]

Continued on next page

Table G.2 – *continued from previous page*

Panel B: Sequential double sorts on $ILMV$ and stock characteristics

		Portfolios of beta						Portfolios of market capitalization					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of $ILMV$	Low	−0.0089	−0.050	−0.29***	−0.35**	−0.90***	−0.89**	−1.02***	−0.49***	−0.039	−0.057	0.0071	1.03***
		[−0.06]	[−0.69]	[−3.68]	[−2.57]	[−2.79]	[−2.12]	[−4.23]	[−2.85]	[−0.31]	[−0.69]	[0.19]	[4.06]
	2	0.19	0.099	−0.12	−0.17	−0.63***	−0.82***	−0.65***	−0.13	0.047	0.032	0.071	0.72***
		[1.31]	[1.55]	[−1.18]	[−1.32]	[−3.65]	[−3.08]	[−3.86]	[−1.18]	[0.43]	[0.32]	[0.87]	[3.41]
	3	0.10	0.23**	0.15	0.11	−0.45***	−0.55**	−0.32**	0.11	0.12*	0.064	0.17*	0.48***
		[0.92]	[2.07]	[1.64]	[1.03]	[−2.73]	[−2.45]	[−2.60]	[1.00]	[1.77]	[0.72]	[1.83]	[2.91]
	4	0.47***	0.50***	0.45***	0.13	−0.14	−0.61***	−0.035	0.40***	0.42***	0.38***	0.23**	0.26
		[4.84]	[5.30]	[3.76]	[1.09]	[−1.02]	[−3.57]	[−0.16]	[3.18]	[3.56]	[3.89]	[2.31]	[0.96]
	High	0.75***	0.77***	0.50***	0.43***	0.30**	−0.45*	0.33*	0.77***	0.65***	0.56***	0.46***	0.13
		[3.78]	[5.70]	[3.14]	[3.43]	[2.25]	[−1.88]	[1.78]	[4.05]	[4.62]	[3.65]	[2.80]	[0.51]

		Portfolios of book-to-market ratio						Portfolios of past return ($R_{(-12,-2)}$)					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of $ILMV$	Low	−0.12	−0.31***	−0.33**	−0.38***	−0.46**	−0.34	−0.99***	−0.11	−0.14**	0.048	−0.40**	0.59
		[−0.64]	[−2.78]	[−2.59]	[−3.07]	[−2.27]	[−1.18]	[−3.05]	[−1.00]	[−2.10]	[0.40]	[−1.99]	[1.40]
	2	−0.12	−0.022	−0.064	−0.28**	−0.13	−0.0098	−0.66***	0.072	0.20*	−0.049	−0.19	0.48
		[−0.98]	[−0.23]	[−0.66]	[−2.23]	[−0.87]	[−0.04]	[−3.48]	[0.62]	[1.93]	[−0.45]	[−1.14]	[1.55]
	3	0.085	−0.053	0.040	−0.043	0.11	0.024	−0.24	0.14	0.045	0.21*	−0.014	0.23
		[0.52]	[−0.50]	[0.45]	[−0.39]	[0.98]	[0.11]	[−1.33]	[1.09]	[0.48]	[1.90]	[−0.08]	[0.72]
	4	0.44***	0.10	0.15	0.38***	0.33**	−0.11	−0.11	0.11	0.47***	0.54***	0.40***	0.51**
		[2.69]	[0.97]	[1.29]	[3.21]	[2.37]	[−0.52]	[−0.65]	[0.82]	[4.09]	[4.53]	[3.57]	[2.29]
	High	0.21	0.30**	0.58***	0.86***	0.80***	0.58**	−0.070	0.59***	0.88***	0.73***	0.63***	0.70***
		[1.43]	[2.34]	[3.54]	[5.21]	[4.54]	[2.49]	[−0.42]	[4.23]	[6.60]	[4.98]	[4.46]	[3.69]

H Three-month *ILM*s and Expected Returns

This section establishes the robustness of our main asset pricing findings to constructing liquidity measures over rolling 3-month windows. We first uncover results similar to those in Table 8 using liquidity measures constructed over rolling 3-month windows. Specifically, $LIQ_{j,m-2}$ averages daily stock j's observations from month $m-4$ through $m-2$. Table H.1 reports that, with a \$2 minimum price requirement, $ILMT$ and $ILMV$ explain the cross-section of stock returns in month m, unlike other liquidity measures. Sample standard deviations for $ILMT$ and $ILMV$ are 0.176 and 0.195, respectively. Thus, a one standard deviation increase in $ILMT$ is associated with estimated monthly liquidity premium of $0.176 \times 1.45\% = 0.255\%$, or 3.06% per year. Similarly, the liquidity premium associated with a one standard deviation increase in $ILMV$ is $0.195 \times 1.60 = 0.312\%$ per month or 3.74% per year.

Table H.1. Liquidity and the Cross-Section of Expected Stock Returns: 3-month *ILM*s. This table reports on the relation between an array of high-frequency liquidity measures and the cross-section of expected stock returns. Equation (2) is estimated using liquidity measures ($LIQ_{j,m-2}$) constructed over 3-month horizons. Control variables include three Fama-French betas ($\beta^{mkt}_{j,m-1}$, $\beta^{hml}_{j,m-1}$, $\beta^{smb}_{j,m-1}$), estimated using weekly observations from the two-year period ending in the final full week of month $m-1$, book-to-market ratio ($\text{BM}_{j,m-1}$), natural log of market capitalization ($\ln(\text{Mcap}_{j,m-1})$), dividend yield ($\text{DYD}_{j,m-1}$), defined as total dividends over the past 12 months divided by the share price at the end of month $m-1$, idiosyncratic volatility ($\text{IdVol}_{j,m-1}$), previous month's return ($RET_{(-1)}$), and preceding return from the prior 11 months ($RET_{(-12,-2)}$). Estimates are from Fama-MacBeth regressions featuring Newey-West corrected standard errors with 6 lags. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$2. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

	InPrIm	QSP	ShrDepth	EFSP	RESP	PIMP	Lambda	AMVST	ROLL	ILLIQ	ILLIQ_OC	BBD	WBBD	ILMT	ILMV
Constant	1.47	0.70	0.71	0.68	0.75	0.71	1.53*	0.96	1.53*	0.92	0.90	1.51*	1.51*	−1.62	−2.40
	[1.17]	[0.76]	[0.79]	[0.73]	[0.84]	[0.79]	[1.71]	[1.12]	[1.70]	[1.06]	[1.00]	[1.73]	[1.75]	[−1.17]	[−1.58]
Liquidity	**0.060**	**0.042**	**−0.00**	**0.11**	**−0.095**	**0.091**	**−0.18****	**−0.038**	**−10.8*****	**−0.041**	**−0.057**	**−0.13**	**−0.19**	**1.45*****	**1.60*****
	[1.28]	**[0.34]**	**[−1.07]**	**[0.64]**	**[−0.77]**	**[0.72]**	**[−2.13]**	**[−0.37]**	**[−4.26]**	**[−1.25]**	**[−0.65]**	**[−0.88]**	**[−0.72]**	**[2.95]**	**[3.26]**
β^{mkt}	−0.039	−0.21	−0.21	−0.21	−0.22	−0.21	−0.23	−0.18	−0.18	−0.22	−0.22	−0.24	−0.24	−0.12	−0.082
	[−0.11]	[−1.04]	[−1.04]	[−1.03]	[−1.05]	[−1.04]	[−1.08]	[−0.89]	[−0.86]	[−1.05]	[−1.06]	[−1.00]	[−0.99]	[−0.62]	[−0.44]
β^{hml}	−0.10	−0.13	−0.13	−0.13	−0.13	−0.13	−0.13	−0.12	−0.12	−0.13	−0.13	−0.10	−0.10	−0.14	−0.16
	[−0.69]	[−1.07]	[−1.06]	[−1.07]	[−1.06]	[−1.06]	[−1.05]	[−0.97]	[−1.03]	[−1.06]	[−1.06]	[−0.72]	[−0.73]	[−1.19]	[−1.27]
β^{smb}	0.12	0.039	0.037	0.039	0.034	0.036	0.015	0.048	0.044	0.023	0.024	0.022	0.022	0.080	0.093
	[1.27]	[0.53]	[0.50]	[0.53]	[0.47]	[0.49]	[0.20]	[0.65]	[0.60]	[0.31]	[0.32]	[0.25]	[0.25]	[1.12]	[1.31]
BM	0.19	−0.026	−0.026	−0.026	−0.027	−0.027	−0.025	0.00040	0.0057	−0.0095	−0.0100	0.026	0.027	−0.029	−0.027
	[1.43]	[−0.54]	[−0.53]	[−0.53]	[−0.56]	[−0.56]	[−0.45]	[0.01]	[0.12]	[−0.19]	[−0.20]	[0.32]	[0.33]	[−0.59]	[−0.55]
ln(Mcap)	0.0010	0.036	0.036	0.037	0.034	0.036	−0.00043	0.023	−0.00017	0.026	0.027	0.0028	0.0028	0.12**	0.15**
	[0.02]	[0.96]	[0.99]	[0.98]	[0.93]	[0.98]	[−0.01]	[0.65]	[−0.00]	[0.74]	[0.74]	[0.08]	[0.08]	[2.24]	[2.54]
DYD	0.34	−0.096	−0.099	−0.091	−0.10	−0.10	−0.034	−0.067	−0.092	−0.065	−0.084	0.12	0.12	−0.14	−0.14
	[0.31]	[−0.17]	[−0.17]	[−0.16]	[−0.18]	[−0.18]	[−0.06]	[−0.12]	[−0.16]	[−0.11]	[−0.15]	[0.18]	[0.18]	[−0.26]	[−0.25]
Id. Vol.	−0.16**	−0.23***	−0.23***	−0.23***	−0.23***	−0.23***	−0.22***	−0.23***	−0.22***	−0.23***	−0.23***	−0.22***	−0.23***	−0.21***	−0.20***
	[−2.57]	[−4.66]	[−4.68]	[−4.66]	[−4.64]	[−4.65]	[−4.43]	[−4.73]	[−4.47]	[−4.51]	[−4.37]	[−3.82]	[−3.82]	[−4.44]	[−4.31]
RET_{-1}	−0.84	−0.33	−0.34	−0.34	−0.33	−0.32	−0.29	−0.34	−0.38	−0.35	−0.34	−0.43	−0.43	−0.41	−0.46
	[−1.16]	[−0.69]	[−0.70]	[−0.70]	[−0.68]	[−0.67]	[−0.61]	[−0.71]	[−0.80]	[−0.72]	[−0.70]	[−0.80]	[−0.80]	[−0.86]	[−0.96]
$RET_{(-12,-2)}$	0.37*	0.21	0.21	0.21	0.21	0.21	0.18	0.21	0.21	0.21	0.21	0.21	0.21	0.28*	0.29*
	[1.96]	[1.35]	[1.34]	[1.35]	[1.35]	[1.35]	[1.12]	[1.39]	[1.35]	[1.35]	[1.30]	[1.07]	[1.07]	[1.71]	[1.81]
Observations	131,828†	327,842	327,842	327,842	327,842	327,842	332,943	337,181	337,185	334,134††	334,134††	271,641†††	271,641†††	327,842	327,842

† The number of observations reflects the largest sample available in ANcerno data from 2010–2014.

†† The number of observations reflects the largest sample available for ILLIQ and ILLIQ_OC.

††† The number of observations reflects the largest sample available for BBD and WBBD from 2010–2017.

Second, we present results from various robustness tests when our institutional liquidity measures are constructed over 3-month rolling windows. Table H.2 reports results similar to those in Table 9 using ILMs constructed over 3-month rolling windows. While our conclusions otherwise remain unchanged, the ILM coefficients in value-weighted regressions do become insignificant.

Table H.2. The Cross-Section of Expected Stock Returns and ILM: Robustness Tests. This table reports on the robustness of the relation between between our institutional liquidity measures and the cross-section of expected stock returns. Equation (2) is estimated using institutional liquidity measures ($LIQ_{j,m-2}$) constructed over 3-month horizons. Control variables include three-factor Fama-French betas ($\beta^{mkt}_{j,m-1}$, $\beta^{hml}_{j,m-1}$, $\beta^{smb}_{j,m-1}$), estimated using weekly observations from the two-year period ending in the final full week of month $m-1$, book-to-market ratio ($\text{BM}_{j,m-1}$), natural log of market capitalization ($\ln(\text{Mcap}_{j,m-1})$), dividend yield ($\text{DYD}_{j,m-1}$), defined as total dividends over the past 12 months divided by the share price at the end of month $m-1$, idiosyncratic volatility ($\text{IdVol}_{j,m-1}$), previous month's return ($RET_{(-1)}$), and preceding return from the prior 11 months ($RET_{(-12,-2)}$). Panel A reports on the robustness of the results to (1) estimating coefficients using panel regressions with date and stock fixed effects and date-stock double-clustered standard errors, (2) weighting observations (by size or according to Asparouhova et al. 2010) to correct for microstructure noise, (3) excluding firms with the smallest 20% market capitalization, (4) excluding stocks in the bottom 10% of the ratio of sub-penny volume in total volume; and (5) excluding stocks in the top or bottom 10% of the respective ILM. Stocks whose previous month-end's closing price is below $p_{min} \in \{\$1, \$2, \$5\}$ are excluded. Panel B reports on the robustness of the estimates in equation (2) to listing exchange. Observations are weighted according to Asparouhova et al. (2010) after excluding stocks whose previous month-end's closing price is below \$1 and stocks falling in the bottom 10% of the ratio of sub-penny volume in total volume. Estimates are from Fama-MacBeth regressions that have Newey-West corrected standard errors with 6 lags. The sample includes NMS common shares from January 2010 to December 2019. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Robustness to estimation method and sample selection

Robustness specification	$ILMT$			$ILMV$		
	Price > \$1	Price > \$2	Price > \$5	Price > \$1	Price > \$2	Price > \$5
Panel regressions + stock & date FEs	1.77**	1.56**	0.61	2.24***	1.89***	1.04*
+ double-clustered S.E.	[2.45]	[2.24]	[0.90]	[3.33]	[2.99]	[1.78]
Asparouhova et al. (2010)	1.54***	1.43***	0.85*	1.74***	1.57***	1.14**
	[2.77]	[2.77]	[1.96]	[3.10]	[3.00]	[2.60]
Asparouhova et al. (2010)	1.21**	1.15**	0.83*	1.43***	1.36***	1.10**
+ top 80% market capitalization	[2.46]	[2.40]	[1.87]	[2.76]	[2.77]	[2.50]
Asparouhova et al. (2010)	1.68***	1.62***	1.07**	1.90***	1.77***	1.37***
+ low sub-penny volume stocks excluded	[2.96]	[3.02]	[2.40]	[3.32]	[3.28]	[3.03]
Size-weighted estimation	1.16	1.18	1.20	0.24	0.24	0.22
	[1.52]	[1.54]	[1.51]	[0.39]	[0.38]	[0.34]
Stocks in top and bottom	3.22***	2.86***	2.03***	2.35***	2.21***	1.88***
10% of ILM excluded	[3.51]	[3.78]	[3.32]	[3.18]	[3.32]	[3.38]

Panel B: Robustness to estimation by listing exchange

	NYSE/AMEX	NASDAQ	NYSE/AMEX	NASDAQ
Asparouhova et al. (2010)	0.81	1.48**	1.35**	1.61***
+ Price > \$1	[1.35]	[2.45]	[2.13]	[2.81]
Asparouhova et al. (2010) + Price > \$1	1.04	1.57**	1.59**	1.71***
+ low sub-penny volume stocks excluded	[1.65]	[2.58]	[2.43]	[2.97]

Third, we report three-factor alphas for long-short trading strategies conditional on ILMs constructed over 3-month ($m-4$ to $m-2$) rolling windows. Table H.3 presents results similar to

those in Table 10. Panel A reports that equal-weighed long-short strategies conditional on 3-month ILMs are associated with monthly three-factor alphas that range from 0.82% to 1.1% depending on minimum share price requirements of \$1, \$2, and \$5. Panel B reports three-factor alphas from long-short strategies based on value-weighted returns calculated after removing stocks with the smallest 20% market capitalization. Alphas range from 0.29% to 0.63% per month, which correspond to annualized three-factor alphas of 3.48% and 7.56%. These results confirm the robustness of liquidity premia to constructing ILMs over 3-month rolling windows.

Tables H.4 and H.5 demonstrate the robustness of our double-sort results to the use of ILMs constructed over 3-month ($m-4$ to $m-2$) rolling windows. We find significant liquidity premia in all subsamples (quintiles) of stock characteristics. In contrast, the momentum anomaly becomes insignificant after controlling for institutional liquidity. The value premium is also more salient among less liquid stocks.

Table H.3. *ILM* **Liquidity Alphas: CRSP and NYSE Breakpoints, Equal- and Value-Weighted Returns.** This table presents three-factor alphas conditional on ILM. Panels A, B, and C report results based on NMS-listed common shares using CRSP breakpoints and equally-weighted portfolio returns. Panels D, E, F report results based on NMS-listed common shares, after first removing stocks with the smallest 20% market capitalization in the prior month, using NYSE breakpoints and value-weighted portfolio returns. Stocks in each monthly cross-section are sorted into ten portfolios (deciles) conditional on one ILM. Monthly portfolio returns are averages of monthly stock returns in the portfolio. The time-series features 116 months. The time-series of returns for each portfolio (after subtracting the 1-month Treasury-bill rate) including the long-short portfolio are then regressed on the Fama-French three factors. The resulting intercepts are three-factor alphas. The sample period is from January 2010 to December 2019, excluding stock's whose previous month-end's closing price is below $p_{min} \in \{\$1, \$2, \$5\}$. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: CRSP breakpoints, \$1 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.61***	−0.43***	−0.34***	−0.24***	−0.15	−0.054	0.020	0.22**	0.33**	0.57***	1.18***
	[−3.68]	[−5.13]	[−3.31]	[−2.67]	[−1.51]	[−0.57]	[0.24]	[2.14]	[2.57]	[4.02]	[4.79]
$ILMV$	−0.32**	−0.37***	−0.21**	−0.20**	−0.13	−0.25*	−0.15	0.093	0.28*	0.58***	0.90***
	[−2.34]	[−3.80]	[−2.60]	[−2.22]	[−1.18]	[−1.91]	[−1.21]	[0.79]	[1.96]	[4.11]	[3.97]

Panel B: CRSP breakpoints, \$2 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.54***	−0.38***	−0.32***	−0.17*	−0.058	0.012	0.033	0.31***	0.27**	0.54***	1.09***
	[−3.49]	[−4.78]	[−3.59]	[−1.92]	[−0.65]	[0.16]	[0.49]	[3.05]	[2.53]	[3.94]	[4.41]
$ILMV$	−0.30**	−0.35***	−0.23***	−0.11	−0.041	−0.14	−0.032	0.073	0.29**	0.55***	0.86***
	[−2.27]	[−4.05]	[−2.74]	[−1.55]	[−0.53]	[−1.33]	[−0.37]	[0.67]	[2.18]	[4.23]	[3.94]

Panel C: CRSP breakpoints, \$5 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.45***	−0.26***	−0.15*	−0.068	0.025	0.11*	0.15**	0.37***	0.40***	0.64***	1.09***
	[−3.23]	[−3.59]	[−1.77]	[−0.71]	[0.36]	[1.68]	[2.16]	[3.71]	[4.02]	[5.20]	[4.74]
$ILMV$	−0.28**	−0.24***	−0.17**	−0.032	0.073	0.11	0.075	0.22**	0.42***	0.61***	0.89***
	[−2.19]	[−2.83]	[−2.48]	[−0.39]	[0.96]	[1.33]	[1.04]	[2.25]	[3.75]	[5.02]	[4.24]

Continued on next page

Table H.3 – *continued from previous page*

Panel D: NYSE breakpoints, largest 80% market capitalization, $1 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.15***	0.088	0.16*	0.085	0.11	0.26**	0.13	0.19	0.23**	0.48***	0.63***
	[−2.72]	[1.08]	[1.81]	[1.00]	[1.05]	[2.31]	[1.15]	[1.59]	[2.59]	[4.14]	[5.06]
$ILMV$	−0.078	−0.0055	0.061	0.17**	0.067	0.063	0.24***	0.32***	0.33***	0.29**	0.37***
	[−1.39]	[−0.07]	[0.90]	[2.01]	[0.74]	[0.64]	[3.79]	[5.13]	[2.74]	[2.07]	[2.62]

Panel E: NYSE breakpoints, largest 80% market capitalization, $2 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.16***	0.10	0.16*	0.070	0.11	0.30**	0.12	0.20*	0.25***	0.46***	0.62***
	[−2.86]	[1.31]	[1.85]	[0.83]	[0.97]	[2.57]	[1.11]	[1.77]	[2.71]	[3.97]	[4.95]
$ILMV$	−0.076	−0.017	0.064	0.19**	0.057	0.071	0.24***	0.34***	0.34***	0.27*	0.34**
	[−1.33]	[−0.20]	[0.97]	[2.23]	[0.65]	[0.76]	[3.57]	[5.67]	[3.03]	[1.87]	[2.44]

Panel F: NYSE breakpoints, largest 80% market capitalization, $5 minimum share price

	Liquidity portfolios										
	1	2	3	4	5	6	7	8	9	10	10 − 1
$ILMT$	−0.15***	0.098	0.13*	0.080	0.095	0.28**	0.15	0.14	0.36***	0.44***	0.58***
	[−2.66]	[1.31]	[1.68]	[0.96]	[0.90]	[2.61]	[1.37]	[0.99]	[2.92]	[4.39]	[5.23]
$ILMV$	−0.065	−0.039	0.073	0.18**	0.092	0.065	0.25***	0.33***	0.31***	0.30**	0.36**
	[−1.14]	[−0.46]	[1.08]	[2.18]	[0.98]	[0.70]	[3.27]	[4.64]	[2.86]	[2.02]	[2.54]

Table H.4. Liquidity Alphas: Stock Characteristic and ILM Double-Sorts. This table presents three-factor alphas to ILM using CRSP breakpoints for stock characteristic quintiles. Stocks are sorted into quintiles of characteristic $X \in \{\beta^{mkt}, \text{Mcap}, RET_{(-12,-2)}, \text{BM}, \text{SPVS}\}$. Within each quintile of characteristic X, stocks are further sorted into quintiles of $LIQ \in \{ILMT, ILMV\}$. Monthly 5×5 portfolio returns are equally-weighted averages of monthly stock returns in the portfolio. The time-series of returns for each portfolio (net of 1-month Treasury-bill rate) including the long-short portfolio are then regressed on the Fama-French three factors. The resulting intercepts are three-factor alphas. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$5. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Sequential double sorts on market beta and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of market beta	Low	0.075	0.15	0.34**	0.67***	0.75***	0.68***	−0.054	0.088	0.50***	0.68***	0.77***	0.83***
		[0.42]	[0.97]	[2.33]	[4.95]	[4.31]	[2.88]	[−0.32]	[0.61]	[4.24]	[4.82]	[4.25]	[3.55]
	2	0.056	0.34***	0.50***	0.48***	0.52***	0.46***	0.14	0.34**	0.43***	0.44***	0.55***	0.41**
		[0.57]	[2.87]	[5.06]	[4.03]	[4.17]	[3.04]	[1.30]	[2.60]	[4.60]	[3.98]	[4.23]	[2.49]
	3	−0.11	−0.023	0.23**	0.33***	0.34***	0.45***	−0.18**	0.042	0.15	0.36***	0.39***	0.57***
		[−1.62]	[−0.26]	[2.29]	[3.88]	[3.32]	[3.43]	[−2.42]	[0.41]	[1.33]	[4.82]	[3.43]	[3.81]
	4	−0.17*	−0.19*	0.0089	−0.089	0.13	0.30	−0.31***	−0.19*	−0.059	0.096	0.14	0.45***
		[−1.80]	[−1.80]	[0.09]	[−0.86]	[0.70]	[1.48]	[−2.83]	[−1.80]	[−0.44]	[0.86]	[0.95]	[2.97]
	High	−0.78***	−0.45**	−0.41**	−0.40**	−0.23	0.56*	−0.85***	−0.54***	−0.27	−0.48***	−0.14	0.71***
		[−2.75]	[−2.22]	[−2.24]	[−2.51]	[−1.25]	[1.78]	[−3.10]	[−2.80]	[−1.52]	[−2.70]	[−0.85]	[2.66]

Panel B: Sequential double sorts on market capitalization and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of market capitalization	Low	−0.70***	−0.20	0.49***	0.65***	0.65***	1.35***	−0.93***	0.11	0.36**	0.65***	0.69***	1.62***
		[−3.31]	[−0.89]	[3.80]	[4.64]	[3.39]	[4.83]	[−3.98]	[0.47]	[2.61]	[4.66]	[3.72]	[5.71]
	2	−0.69***	−0.018	0.23*	0.54***	0.50***	1.19***	−0.79***	−0.064	0.31***	0.53***	0.56***	1.34***
		[−3.72]	[−0.16]	[1.86]	[4.39]	[2.80]	[4.47]	[−3.82]	[−0.59]	[2.71]	[4.05]	[3.11]	[4.63]
	3	−0.33**	0.13	0.19**	0.18**	0.35***	0.68***	−0.44***	0.11	0.17	0.31***	0.36***	0.79***
		[−2.36]	[1.23]	[2.08]	[2.09]	[3.01]	[3.81]	[−3.00]	[1.11]	[1.60]	[2.93]	[3.12]	[3.99]
	4	−0.37*	−0.18*	0.011	0.13*	0.23**	0.60**	−0.52**	−0.093	0.013	0.17*	0.26***	0.77***
		[−1.75]	[−1.93]	[0.10]	[1.92]	[2.52]	[2.62]	[−2.36]	[−0.80]	[0.16]	[1.94]	[3.24]	[3.30]
	High	−0.26**	−0.00081	0.021	0.17**	0.27***	0.53***	−0.17	−0.041	0.017	0.19**	0.21**	0.38***
		[−2.48]	[−0.01]	[0.29]	[2.18]	[4.21]	[4.34]	[−1.27]	[−0.57]	[0.21]	[2.27]	[2.19]	[2.94]

Continued on next page

Table H.4 – *continued from previous page*

Panel C: Sequential double sorts on book-to-market ratio and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of book-to-market ratio	Low	−0.14	−0.21*	0.040	0.17	0.20	0.34	−0.33*	−0.21*	0.13	0.26*	0.21	0.54**
		[−0.92]	[−1.75]	[0.34]	[0.99]	[0.97]	[1.45]	[−1.70]	[−1.80]	[1.16]	[1.66]	[1.06]	[2.44]
	2	−0.25*	−0.25**	0.020	0.13	0.13	0.38*	−0.37***	−0.17*	−0.0020	0.059	0.26	0.63***
		[−1.85]	[−2.48]	[0.20]	[1.09]	[0.74]	[1.68]	[−2.75]	[−1.74]	[−0.01]	[0.54]	[1.63]	[2.92]
	3	−0.22*	−0.047	0.071	0.028	0.12	0.34*	−0.23*	−0.11	0.063	0.035	0.20*	0.43**
		[−1.97]	[−0.45]	[0.93]	[0.23]	[0.94]	[1.71]	[−1.73]	[−0.91]	[0.87]	[0.27]	[1.68]	[2.03]
	4	−0.29**	−0.040	0.13	0.36***	0.72***	1.01***	−0.36***	−0.12	0.19*	0.41***	0.77***	1.13***
		[−2.19]	[−0.34]	[0.93]	[2.87]	[4.82]	[4.35]	[−2.68]	[−0.87]	[1.77]	[3.12]	[4.88]	[4.47]
	High	−0.36**	0.037	0.22	0.58***	0.81***	1.17***	−0.48**	0.074	0.25*	0.61***	0.82***	1.29***
		[−2.00]	[0.21]	[1.28]	[3.63]	[5.07]	[4.00]	[−2.19]	[0.45]	[1.70]	[4.19]	[5.44]	[4.06]

Panel D: Sequential double sorts on past 11-month return and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of past return	Low	−0.99***	−0.48**	−0.072	−0.20	−0.069	0.92***	−0.92***	−0.65***	−0.25	0.065	−0.071	0.84**
		[−3.57]	[−2.05]	[−0.32]	[−1.18]	[−0.37]	[2.66]	[−2.85]	[−3.11]	[−1.23]	[0.35]	[−0.44]	[2.30]
	2	−0.11	−0.037	0.17	0.27**	0.50***	0.61***	−0.13	−0.022	0.11	0.34**	0.51***	0.64***
		[−0.88]	[−0.31]	[1.39]	[2.06]	[3.46]	[3.08]	[−0.89]	[−0.17]	[0.91]	[2.57]	[3.52]	[2.79]
	3	−0.037	0.088	0.34***	0.38***	0.95***	0.98***	−0.11	0.21*	0.19*	0.44***	0.98***	1.09***
		[−0.49]	[0.76]	[3.19]	[2.63]	[6.65]	[5.79]	[−1.28]	[1.86]	[1.74]	[3.08]	[6.78]	[6.65]
	4	−0.046	0.15	0.16	0.27**	0.72***	0.77***	−0.12	0.088	0.24**	0.32***	0.73***	0.84***
		[−0.57]	[1.16]	[1.50]	[2.37]	[5.00]	[4.36]	[−1.27]	[0.82]	[1.99]	[2.95]	[4.90]	[4.44]
	High	−0.34*	−0.23	0.028	0.096	0.44***	0.78***	−0.48**	−0.16	−0.087	0.24	0.48***	0.96***
		[−1.68]	[−1.15]	[0.16]	[0.55]	[2.86]	[3.47]	[−2.03]	[−0.89]	[−0.52]	[1.42]	[3.07]	[3.63]

Continued on next page

Table H.4 – *continued from previous page*

Panel E: Sequential double sorts on share of sub-penny trade volume and ILM

		Portfolios of $ILMT$						Portfolios of $ILMV$					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of sub-penny volume	Low	0.078	−0.072	0.29***	0.12	0.46***	0.38**	0.046	−0.0051	0.18**	0.18**	0.48***	0.43**
		[0.78]	[−0.67]	[3.52]	[1.29]	[3.62]	[2.12]	[0.51]	[−0.05]	[2.03]	[2.17]	[3.55]	[2.26]
	2	−0.011	0.041	0.22**	0.23***	0.38***	0.39**	−0.096	0.12	0.19**	0.30***	0.35***	0.45***
		[−0.11]	[0.50]	[2.57]	[3.12]	[3.41]	[2.54]	[−1.07]	[1.12]	[2.28]	[3.66]	[3.16]	[2.93]
	3	−0.13	−0.10	0.055	0.0026	0.54***	0.67***	−0.11	−0.21**	0.0068	0.13	0.55***	0.66***
		[−1.29]	[−1.28]	[0.58]	[0.02]	[4.20]	[4.16]	[−0.88]	[−2.30]	[0.08]	[0.97]	[3.52]	[3.15]
	4	−0.19	−0.034	0.057	0.17	0.61***	0.80***	−0.14	−0.13	−0.012	0.27**	0.64***	0.78***
		[−1.64]	[−0.26]	[0.37]	[1.54]	[3.51]	[3.27]	[−1.12]	[−0.91]	[−0.07]	[2.33]	[4.04]	[3.20]
	High	−1.28***	−0.64***	−0.21	0.43***	0.77***	2.05***	−1.25***	−0.86***	−0.068	0.44***	0.81***	2.05***
		[−5.00]	[−3.82]	[−0.92]	[2.89]	[4.18]	[5.98]	[−4.62]	[−4.66]	[−0.32]	[2.75]	[4.19]	[5.54]

Table H.5. Liquidity Alphas: ILM and Stock Characteristic Double-Sorts. This table presents three-factor alphas associated with ILMs and stock characteristics using CRSP breakpoints. Stocks are sorted into quintiles of $LIQ \in \{ILMT, ILMV\}$ constructed over three-month rolling windows. Within each LIQ quintile, stocks are further sorted into quintiles of characteristic $X \in \{\beta^{mkt}, \text{Mcap}, RET_{(-12,-2)}, \text{BM}\}$. Monthly 5×5 portfolio returns are equally-weighted averages of monthly stock returns in the portfolio. The time-series of returns for each portfolio (after subtracting the 1-month Treasury-bill rate) including the long-short portfolio are then regressed on the Fama-French three factors. The resulting intercepts are three-factor alphas. The sample includes NMS common shares from January 2010 to December 2019, excluding stocks whose previous month-end's closing price is below \$5. The numbers in brackets are t-statistics with ***, **, and * identifying statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Sequential double sorts on $ILMT3$ and stock characteristics

		Portfolios of beta						Portfolios of market capitalization					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of $ILMT$	Low	0.11	0.013	−0.19*	−0.32**	−0.79**	−0.90**	−0.86***	−0.25	−0.096	−0.013	0.027	0.88***
		[0.94]	[0.16]	[−1.80]	[−2.10]	[−2.51]	[−2.39]	[−3.38]	[−1.63]	[−0.72]	[−0.14]	[0.71]	[3.48]
	2	0.061	−0.093	−0.19	−0.35*	−0.83***	−0.89***	−0.74**	−0.33	−0.26	−0.16	0.072	0.81***
		[0.30]	[−0.55]	[−1.21]	[−1.79]	[−3.03]	[−3.18]	[−2.56]	[−1.64]	[−1.64]	[−1.07]	[0.51]	[3.64]
	3	0.036	0.17	−0.12	−0.38	−0.64	−0.68**	−0.54	−0.35	0.10	−0.19	0.034	0.57**
		[0.14]	[0.75]	[−0.54]	[−1.36]	[−1.62]	[−2.36]	[−1.39]	[−1.37]	[0.41]	[−0.81]	[0.15]	[2.18]
	4	0.016	0.30	0.085	−0.28	−0.77**	−0.79***	−0.62*	−0.071	0.027	0.068	−0.062	0.56**
		[0.06]	[1.20]	[0.35]	[−1.09]	[−2.11]	[−3.55]	[−1.75]	[−0.24]	[0.11]	[0.26]	[−0.25]	[2.18]
	High	0.51*	0.46	0.30	0.17	−0.13	−0.64**	−0.022	0.44	0.35	0.28	0.26	0.28
		[1.69]	[1.61]	[0.99]	[0.60]	[−0.39]	[−2.54]	[−0.07]	[1.30]	[1.29]	[0.92]	[0.88]	[1.14]

		Portfolios of book-to-market ratio						Portfolios of past return ($R_{(-12,-2)}$)					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of $ILMT$	Low	−0.14	−0.26**	−0.26*	−0.17	−0.36**	−0.22	−0.77***	−0.0068	−0.052	−0.074	−0.28	0.50
		[−0.79]	[−2.41]	[−1.84]	[−1.41]	[−2.31]	[−1.02]	[−2.99]	[−0.05]	[−0.62]	[−0.63]	[−1.33]	[1.51]
	2	−0.10	−0.29	−0.35*	−0.31	−0.36*	−0.26	−0.79***	−0.13	0.046	−0.061	−0.48	0.31
		[−0.53]	[−1.60]	[−1.89]	[−1.52]	[−1.75]	[−1.31]	[−3.06]	[−0.68]	[0.29]	[−0.30]	[−1.64]	[0.86]
	3	−0.049	−0.17	−0.17	−0.34	−0.22	−0.17	−0.54*	−0.069	0.0096	−0.099	−0.25	0.28
		[−0.17]	[−0.78]	[−0.63]	[−1.09]	[−0.68]	[−0.81]	[−1.67]	[−0.29]	[0.04]	[−0.37]	[−0.68]	[0.85]
	4	−0.14	−0.15	−0.20	−0.066	−0.096	0.040	−0.55*	−0.13	0.099	0.044	−0.12	0.43
		[−0.46]	[−0.63]	[−0.79]	[−0.22]	[−0.32]	[0.21]	[−1.82]	[−0.54]	[0.36]	[0.15]	[−0.39]	[1.61]
	High	0.034	−0.0019	0.33	0.51	0.43	0.40	−0.42	0.19	0.63**	0.53*	0.38	0.80***
		[0.11]	[−0.01]	[1.06]	[1.60]	[1.13]	[1.29]	[−1.36]	[0.63]	[2.27]	[1.78]	[1.22]	[4.68]

Continued on next page

Table H.5 – *continued from previous page*

Panel B: Sequential double sorts on $ILMV3$ and stock characteristics

		Portfolios of beta						Portfolios of market capitalization					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of $ILMV$	Low	−0.024	−0.070	−0.31***	−0.39***	−0.90***	−0.87**	−1.08***	−0.55***	−0.077	−0.0088	0.011	1.09***
		[−0.17]	[−0.95]	[−3.10]	[−2.74]	[−2.91]	[−2.22]	[−4.50]	[−2.88]	[−0.64]	[−0.10]	[0.33]	[4.41]
	2	0.077	0.050	−0.28*	−0.50**	−0.76***	−0.84***	−0.81***	−0.40*	−0.18	−0.085	0.051	0.87***
		[0.37]	[0.30]	[−1.81]	[−2.54]	[−2.95]	[−2.96]	[−3.00]	[−1.92]	[−1.19]	[−0.54]	[0.33]	[3.93]
	3	−0.012	0.076	−0.093	−0.20	−0.81**	−0.79***	−0.70*	−0.20	−0.025	−0.12	0.0046	0.70***
		[−0.05]	[0.32]	[−0.44]	[−0.66]	[−2.25]	[−2.85]	[−1.87]	[−0.73]	[−0.11]	[−0.52]	[0.02]	[3.21]
	4	0.18	0.31	0.060	−0.12	−0.51	−0.69***	−0.42	0.12	0.11	0.17	−0.064	0.35
		[0.69]	[1.26]	[0.23]	[−0.44]	[−1.45]	[−3.66]	[−1.06]	[0.42]	[0.46]	[0.64]	[−0.25]	[1.21]
	High	0.52*	0.43	0.24	0.21	−0.067	−0.59**	0.012	0.40	0.38	0.31	0.23	0.22
		[1.73]	[1.57]	[0.81]	[0.77]	[−0.21]	[−2.53]	[0.03]	[1.26]	[1.50]	[1.01]	[0.77]	[0.89]

		Portfolios of book-to-market ratio						Portfolios of past return ($R_{(-12,-2)}$)					
		Low	2	3	4	High	High−Low	Low	2	3	4	High	High−Low
Portfolios of $ILMV$	Low	−0.28	−0.34***	−0.30*	−0.23*	−0.54***	−0.25	−1.00***	−0.061	−0.040	−0.077	−0.52**	0.49
		[−1.42]	[−3.09]	[−1.95]	[−1.76]	[−2.71]	[−0.88]	[−3.20]	[−0.47]	[−0.54]	[−0.63]	[−2.15]	[1.14]
	2	0.010	−0.40**	−0.26	−0.45**	−0.32	−0.33*	−0.81***	−0.15	−0.030	−0.11	−0.33	0.48
		[0.06]	[−2.33]	[−1.39]	[−2.33]	[−1.47]	[−1.70]	[−3.46]	[−0.79]	[−0.18]	[−0.57]	[−1.25]	[1.59]
	3	0.039	−0.15	−0.23	−0.45	−0.24	−0.28	−0.62**	−0.023	−0.11	−0.027	−0.26	0.36
		[0.14]	[−0.71]	[−0.85]	[−1.54]	[−0.82]	[−1.42]	[−2.05]	[−0.10]	[−0.47]	[−0.10]	[−0.74]	[1.30]
	4	0.035	−0.13	−0.16	0.047	0.13	0.092	−0.24	−0.055	0.13	0.093	−0.016	0.22
		[0.11]	[−0.56]	[−0.60]	[0.15]	[0.42]	[0.54]	[−0.76]	[−0.22]	[0.50]	[0.32]	[−0.05]	[0.73]
	High	0.047	0.029	0.37	0.47	0.41	0.36	−0.45	0.23	0.68**	0.50*	0.38	0.83***
		[0.19]	[0.12]	[1.19]	[1.49]	[1.14]	[1.41]	[−1.48]	[0.79]	[2.47]	[1.70]	[1.30]	[5.17]

Would Order-By-Order Auctions Be Competitive?*

Thomas Ernst†, Chester Spatt‡, and Jian Sun§

March 12, 2023

Abstract

Retail trading flow is segregated from non-retail flow in U.S. equities, consistent with market segmentation. We model theoretically two methods of executing segregated retail trades: a) broker's routing, whereby brokers evaluate and allocate orders based on each market maker's aggregate performance, and b) order-by-order auctions, where market makers bid on each individual order, a market structure recently proposed by the SEC. We find that order-by-order auctions improve allocative efficiency among market makers, but a winner's curse problem in the auction can reduce retail investor welfare, particularly at times of limited liquidity. Introducing more market participants who compete for retail orders can harm both total efficiency and investor welfare if these new market participants have superior information compared to incumbent wholesalers. Empirical analysis of Retail Liquidity Programs (RLP) currently offered by exchanges shows that these programs behave similar to order-by-order auctions in our model.

*For helpful comments and feedback we are thankful to: Yashar Barardehi, Robert Battalio, Svetlana Bryzgalova, Pete Kyle, Dmitry Livdan, Mark Loewenstein and Bart Zhou Yueshen, along with seminar participants at the Univeristy of Maryland, Arizona State University, and numerous anonymous industry participants.

†University of Maryland, Robert H. Smith School of Business: ternst@umd.edu

‡Carnegie Mellon University, Tepper School of Business: cspatt@andrew.cmu.edu

§Singapore Management University, Lee Kong Chian School of Business: jiansun@smu.edu.sg

I. Introduction

Retail order flow in U.S. equities is segregated, with retail brokers routing almost all their retail customer orders directly to market makers. These market makers assume a best execution obligation once they receive the order, whether they privately internalize trades off-exchange, or fill the retail orders from liquidity sourced from other sources, like exchanges or alternative trading systems (such as "dark pools"). Retail trades are attractive to market makers, either due to lower adverse selection, as in Battalio and Holden (2001), or due to their trades being less correlated, as in Baldauf, Mollner, and Yueshen (2022). In both cases market makers are willing to give retail investors better prices than the exchanges due to greater ability to segregate orders. Recently, the SEC has proposed a change in market structure with the goal of potentially increasing competition among market makers.[1] While previous academic work has explored *whether* retail flow should be segregated and its value, once segregated, the question of *how* retail flow should be executed is comparatively unexplored.

We model and evaluate empirically two distinct methods of executing segregated retail trades: broker's routing and an order-by-order auction. Our broker's routing model closely resembles the current market structure, with retail brokers determining where to route each order to maximize execution quality. While retail brokers use recent past competing market maker performance to inform routing decisions, they do not communicate with the market maker prior to routing each individual order. Our order-by-order auction models the SEC's proposed Rule 615, which would mandate auctions for retail trades Securities and Exchange Commission (2022). These auctions would only be available for retail market orders, but any market participant could bid on each individual order, yet no one would be required to bid on any given order.

We evaluate both models with a focus on inventory cost and competition. In our model, a broker receives an order from a retail investor and chooses one market maker to execute the order. Executing the order incurs (marginal) inventory costs for market makers. We assume that each market maker i has a private liquidity signal y_i, and that inventory cost is affected by both the market maker's private signal and the average signal of all market makers. Intuitively, each private

[1] In remarks before the SIFMA Annual meeting, SEC Chair Gensler stated "I've also sought recommendations around how to instill greater competition for retail market orders on an order-by-order basis, through auctions. With greater competition, more market participants would have access to these retail market orders." Gensler (2022).

signal can be thought of as the inventory position of market maker i, with a market maker's willingness to trade depending both on his private inventory and the aggregate liquidity of all market makers. To obtain the order, market participants submit their spreads simultaneously to the broker, and the one with the lowest bid can obtain the order. The key difference between broker's routing and order-by-order auctions is the market participants' information set when they submit their spreads. We solve the market equilibrium under both trading mechanisms and then identify differences in welfare distribution, inventory-management, and order allocation efficiency that arise under each of the two market structures.

In the broker's routing setting, market makers can only observe a noisy version of their liquidity signals when submitting their spreads. The symmetric equilibrium bid (spread) strategy is monotone in the noisy signal, which may be different from the true liquidity signal that determines inventory cost. The broker routes the order to the market maker who submits the lowest spread. This delivers a highly competitive outcome–with relatively low expected market maker profits–because market makers' bidding strategies rely less on their signals, which are just noisy versions of their true private liquidity signals. This closely mirrors the current system of order routing in equities, where market makers agree to accept order flow from brokers, but there is no pre-trade communication on individual orders. Brokers route to a market maker, and the market maker must accept the order. In practice, evaluation of trades is done on a periodic (e.g., daily, weekly, or monthly) basis, and market makers compete on the aggregate execution quality they deliver, rather than bidding against each other on each individual order. This is consistent with our setting that when they compete, they only observe noisy information about their true liquidity/shocks when receiving the order, and the spread is less sensitive to their true liquidity cost/positions. The broker's routing setting delivers strong competition, but the lack of communication on any specific trade means that a trade may be routed to a market maker who has observed high ex-post inventory cost, leading to inefficient order allocation and inventory management.

In the order-by-order auction model, in contrast, brokers bid after observing their true liquidity signals. This is motivated by the proposal that all retail orders have to be auctioned order by order, and thus when market makers compete for retail orders, they already have accurate information about the marginal inventory cost of executing the order. In the auction, market makers' symmetric equilibrium bid (spread) is increasing their private signals y_i, and thus in

equilibrium, the participant with the lowest realized inventory cost will always win the auction with the most aggressive bid, leading to the first best allocative efficiency. The common-value nature of the auction, however, creates a winner's curse problem; whichever participant wins the auction learns that all other participants had higher signals of cost. Consequently, market participants bid conservatively in the auction, and thus they will earn a positive expected profit from the auction because of the strategic concern. We show that this effect is more severe in order-by-order auctions (compared to broker's routing) when competition happens after market makers observe more precious liquidity signals. This implies that when the common-value component in the inventory cost is more important, the welfare effect from the winner's curse is more significant, and thus investor's welfare is more likely to be lower under order-by-order auctions compared to that under broker's routing.

We then examine further the welfare comparison between order-by-order auctions and broker's routing. The welfare of investors can be lower in the order-by-order auction setting at times of limited liquidity. Intuitively, market makers compete after observing their signals. When their signals are more precise about their true liquidity signals, they are more informationally heterogeneous, and their bidding strategies will rely more on their observed signals. Limited liquidity leads to less pressure from auction competitors, less aggressive bids, and larger profits for trading against retail orders. While order-by-order auctions have higher allocative efficiency than broker's routing, order-by-order auctions have *less* competition than broker's routing.

We then extend our baseline model to include institutional traders, as a key objective of the SEC proposal is to enable institutional traders to trade directly with retail investors in auctions. While institutional traders can increase the number of bidders in an auction, they also have superior information about the fundamental value of the asset. Incumbent wholesalers, who have an inventory signal but have no information about the fundamental value, respond by bidding more cautiously in the auction. As a result, the overall welfare of retail investors can further decline in the switch to order-by-order auctions. Moreover, we also find an interesting market segmentation result due to asymmetric information. When information asymmetries between institutional investors and incumbent wholesalers are sufficiently severe, only institutional investors will effectively compete for high-quality (low-cost) orders, while all market participants compete for low-quality (high-cost) orders. This leads to heterogeneous impacts of switching to order-by-order auctions on orders with

different qualities.

We then examine impacts of market design in the cross-section of liquidity. Under broker's routing, a broker can evaluate a wholesaler on the performance across all orders, including different sizes, or stocks of different liquidity. This enables cross-subsidization, where wholesalers may make losses trading small stocks, compensated by profits trading large stocks. Switching to order-by-order auctions can substantially decrease market maker incentives to trade small stocks. As a result, the drop in small-stock liquidity, as well as retail investor welfare, can be particularly precipitous in smaller, less liquid stocks.

While order-by-order auctions only exist as an SEC proposal, we identify a currently-existing close empirical analogue of Retail Liquidity Programs (RLP). Exchange RLP's allow market participants to provide liquidity to retail orders at will by posting hidden limit orders which are only accessible by retail investor orders. When there is at least one round lot (100 shares) of RLP interest, exchanges will disseminate an RPI Flag in the market data indicating the presence of RLP liquidity, though not revealing the exact size or price of the order. If multiple participants post in an RLP, the participant with the most aggressively priced order will have first priority for any incoming retail market order, mirroring the potential competitiveness and allocative efficiency of an order-by-order auction. Unlike the broker's routing system, where market makers must accept any flow the broker routes to them, posting limit orders in a RLP is entirely voluntary: there may be many market participants posting limit orders, or none at all.

Five exchanges currently offer Retail Liquidity Programs. RLPs have times with high levels of market participant interest, with at least one-sided interest quoted for 20% of the day in Russell 1000 stocks, and over 50% of the day for our sample of liquid ETFs. The trading volume executed in RLPs, however, is small, averaging less than 0.3% of total trading volume, despite exchange trading fees being substantially reduced for trades in the program.

Volume in RLPs is higher in stocks that are tick constrained, consistent with RLPs being particularly effective in more liquid stocks. Volume in RLPs increases during periods of higher volatility, while volume for off-exchange sub-penny trading decreases. Under the pecking order theory of Menkveld, Yueshen, and Zhu (2017), RLPs would rank high in the pecking order of venues, with market makers already sourcing liquidity from them to the extent that liquidity is available. Order imbalances during times when the RPI Flag is active are much lower than order

imbalances during the times when the RPI Flag is not active, providing further support for the volatility-sensitive nature of voluntary market participant participation in RLP programs. Price impacts of trades in RLP programs are more sensitive to volatility; when volatility is 1% higher, exchange sub-penny trades have a price impact ten basis points higher, while off-exchange trades have a price impact of only two basis points higher. Consistent with our model, market makers appear to consistently provide stable execution quality, while the RLPs function like order-by-order auctions in our model, with much more variation in outcomes.

When the RPI Flag is active, mid-quote trading is more common off-exchange as well as on-exchange. The distribution of sub-penny trades is roughly similar, with most of the shift in volume coming from at-quote trading switching to mid-quote trading. These mid-quote trades could come from either retail trading or non-retail hidden liquidity trades; only sub-penny on-exchange trades are anonymously identifiable as having a retail participant. Quoted bid-ask spreads tend to be more stable when the retail flag is active, consistent with RLPs supplying liquidity during times of high liquidity. When the retail flag is not active, quoted bid-ask spreads tend to be much wider before and after trades.

The SEC notion of an order-by-order auction seeks to "instill greater competition for retail market orders." Under the current system of broker's routing, each order is sent to a single market maker with no pre-trade communication, and competition is measured by aggregate execution quality. Switching to an order-by-order auction offers a tempting increase in allocative efficiency, as the market participant with the most optimistic signal always wins an auction. But this comes with a drawback, as the participant who wins by outbidding all competitors with less optimistic signals suffers the auction winner's curse. Participants scale back their bids, and obtain increased welfare in the order-by-order system. Retail investor welfare can decrease in the switch to order-by-order trading, particularly for volatile stocks and stocks with few competing liquidity providers.

II. Literature and Contribution

Several prior papers argue retail segmentation is optimal as a market design. Battalio and Holden (2001) argue retail investors have lower adverse selection, while Baldauf et al. (2022) argue retail investors are less correlated. Under both cases, it is optimal to segregate retail flow, but

different mechanisms for segregating retail flow are not explored. Motivated by the recent SEC call for order-by-order competition, our paper provides a theoretical analysis into two possible methods of executing retail trades: the current system of broker's routing, and a hypothetical order-by-order system. We show that the proposed order-by-order system would potentially increase allocative efficiency, but decreases retail investor welfare in less liquid stocks.

Several studies examine how retail participants themselves impact market liquidity. Eaton, Green, Roseman, and Wu (2022), for example, highlight how retail traders can increase order imbalances and volatility, while Parlour and Rajan (2003) argue segmentation decreases consumer welfare, as it leads to a subsidization of retail limit orders. We do not explore the market vs. limit order trade-off, but instead focus on retail marketable orders. Under the SEC vision, retail marketable orders would primarily interact with market maker and non-market maker limit orders through an order-by-order system similar to the current Retail Liquidity Programs. We empirically analyze these RLP programs and find that they have low liquidity in small stocks and at volatile times, matching our model prediction of how order-by-order systems would function.

One possible analogue to order-by-order trading exists in the option markets, where a considerable share of volume executes in auctions. Bryzgalova, Pavlova, and Sikorskaya (2022) show that these auctions are correlated with retail trading measures, while Ernst and Spatt (2022) present empirical analysis of specific rules, such as a price-match guarantee and out-sized allocation, which prevent competition in option auctions. Hendershott, Khan, and Riordan (2022) present a model and empirical evidence that auctions in option markets are imperfectly competitive.

Several recent studies have looked at payment for order flow and segmentation. Comerton-Forde, Malinova, and Park (2018) show that a Canadian trade-at rule which decreases retail segmentation leads to liquidity improvements to lit markets but harms retail trade execution quality. Hu and Murphy (2022), Jain, Mishra, O'Donoghue, and Zhao (2020), and Schwarz, Barber, Huang, Jorion, and Odean (2022) all explore variation in execution quality among brokers. Market makers can offer two possible forms of superior prices: PFOF (payments from market makers to brokers) and price improvement (payments from market makers directly to retail customers). Brokers may or may not pass on the total extent of PFOF revenue back to their customers in the form of lower commissions, as documented in Battalio, Jennings, and Selway (2001), while Schwarz et al. (2022) and Battalio and Jennings (2022) highlight that brokers prioritize execution quality even along dimensions not

reflected in SEC 605 reports. In our welfare analysis, we assume that market makers compete solely on price improvement, akin to PFOF being either zero or entirely passed through to retail investors. Our focus is not on the revenue split of PFOF vs. price improvement, but rather what form of market design delivers overall superior or inferior welfare to final retail investors.

Liquidity varies considerably in the cross-section of stocks. Corwin and Coughenour (2008) argue specialists allocate attention to more liquid stocks during times of market stress, while Foley, Liu, Malinova, Park, and Shkilko (2020) show how tying DMM assignments in large and small stocks can lead to substantial increases in liquidity for small stocks with little to no observed harm for large stocks. In an extension to our model, we show how broker's routing can enable a similar cross-subsidization, which is not possible under order-by-order auctions.

Previous studies (Bernhardt and Hughson (1997) and Biais, Martimort, and Rochet (2000)) show that market makers can earn positive profits when competing for orders. Bernhardt and Hughson (1997) emphasize the importance of order splitting in the duopoly case, while Biais et al. (2000) study common-value auctions where multiple market makers compete for an informed order. In both papers, the key friction is the asymmetric information from the liquidity demand side, which refers to informed traders. Our study also predicts that market makers will earn positive profits in both the broker's routing and order-by-order auction settings. However, in contrast to the previous studies, there is no asymmetric information from the liquidity demand side in our model since retail orders are typically uninformed. In our model, market makers receive private signals about their inventory position, which weakens competition and ensures positive profits in equilibrium. Additionally, we extend our study to institutional traders who can privately obtain signals about asset quality and compete for order flows, as suggested by the SEC. We show that the additional adverse selection on the liquidity supply side may exacerbate market inefficiency, leading to a novel market segmentation prediction.

Our empirical analysis focuses on Retail Liquidity Programs (RLP) offered by several exchanges. Jain, Linna, and McInish (2021) provide an overview of the NYSE Retail Liquidity Program in 2015. Five exchanges now operate RLPs, and we analyze current RLP data through the lens of learning about potential execution quality under the SEC's proposed order-by-order auctions. RLPs provide a competitive process for both traditional market makers and institutional investors to enter limit orders which offer potential price improvement to retail trades, but empirically suffer

from the same winner's curse problem we identify in our theoretical model.

III. Model

The model consists of only two dates, time 0 and time 1, and there is no discounting. There are three types of market players: a (retail) investor, a broker, and $N > 3$ ex-ante identical market makers indexed by $i \in \{1, 2, ...N\}$. Our focus is the strategic interactions among market makers, so we abstract away from agency problems between the investor and the broker, and assume that the broker's objective is to maximize the investor's welfare, which in our model is equivalent to minimizing the spread.

At time 0, the broker receives a one unit sell order from the investor, and sends it to a market maker to execute the order by the end of time 0.[2] We assume that the retail investor is trading only for liquidity reasons, so there is no information about asset value contained in the direction of the order. If market maker i executes the order, it has to hold the position until time 1 which incurs (marginal) inventory cost ζ_i. The structure of ζ_i is specified later in this section. Let s_i be the half bid-ask spread offered by market maker i, then the profit that market maker i receives at time 1 is

$$s_i - \zeta_i.$$

We consider a tractable framework with linear equilibrium in the literature of common-value auctions (Klemperer (1999), Menezes and Monteiro (2004)). At time 0, each market maker i receives an i.i.d private liquidity shock y_i. For simplicity, we assume that y_i is drawn from an uniform distribution $U[-\frac{1}{2}, \frac{1}{2}]$. The inventory cost ζ_i has the following structure

$$\zeta_i = c_0 + c_1 \frac{1}{N} \sum_{j=1}^{N} y_j + c_2 y_i,$$

where c_0, c_1 and c_2 are positive constants. Since each market maker can only observe his own liquidity shock, the inventory cost ζ_i is not fully observed by market maker i. The cost function consists of three components. The first term c_0 is the unconditional expected inventory cost of

[2]The direction of the order does not change our results.

executing the order, which is the same for all market makers. The second term

$$c_1 \frac{1}{N} \sum_{j=1}^{N} y_j$$

represents ζ_i's exposure to the aggregate liquidity shock $\frac{1}{N}\sum_{j=1}^{N} y_j$. When c_1 is higher, the inventory cost of executing the order is more sensitive to the aggregate liquidity shock. The third term

$$c_2 y_i$$

represents ζ_i's exposure to the individual liquidity shock y_i, and the coefficient c_2 measures the sensitivity. In our model, the coefficients (c_0, c_1, c_2) are exogenous, and are determined by stock characteristics. For example, a stock that is about to announce earnings may have a very high c_1, with market makers very concerned about aggregate inventory imbalances. In contrast, a tick-constrained stock with low informational asymmetries may have a very low c_1 value, with market makers not very concerned about aggregate inventories.

A. Order-by-order Auction

First, we consider a hypothetical order-by-order auction mechanism. We model the order-by-order auction as a common-value auction. In order-by-order auctions, each market maker i submits the spread s_i after privately observing the realization of signal y_i at time 0, and thus it can choose its spread strategy according to its assessment of inventory cost. The broker observes spreads offered by all market makers, and sends the order to the winner with the lowest spread at the end of time 0. If more than one market makers submit the lowest spread, then the winner is chosen randomly among those who submit the lowest spread. At time 1, all players collect their payoffs. We focus on symmetric equilibria such that all market makers choose the same strategy.

Intuitively, when observing a higher signal realization y_i, the inventory cost ζ_i tends to be larger for market maker i, and thus it will submit a higher spread s_i. We conjecture (and verify later) that there exists a linear symmetric equilibria where all market makers choose the same strategy $s_i(y) = s(y)$ where

$$s = k_0 + k_1 y.$$

We solve the equilibrium using the standard mechanism design approach. Heuristically, suppose all market makers except market maker i follow the aforementioned equilibrium strategy. We consider market maker i's expected profit $U(z, y)$ where y is the private signal observed by market maker i, and

$$\tilde{s} = k_0 + k_1 z$$

is the spread that market maker i submits to the broker. In equilibrium, we must have

$$\left.\frac{\partial U(z, y)}{\partial z}\right|_{z=y} = 0.$$

The following proposition summarizes our results.

Proposition 1. *In the model of order-by-order auctions, there exists a linear symmetric equilibrium in which the spread submitted by market maker $i \in \{1, 2, ...N\}$ is*

$$s_i(y_i) = k_0 + k_1 y_i,$$

where

$$k_0 = c_0 + \frac{c_1}{4N}\left(N - 1 + \frac{2}{N}\right) + \frac{c_2}{2N}$$

and

$$k_1 = \frac{N-1}{N}\left(\frac{c_1}{2}\frac{N+2}{N} + c_2\right).$$

Proof. See Appendix. □

First, as we discussed earlier, the equilibrium strategy $s_i(y_i)$ is increasing in y_i with sensitivity

$$k_1 = \frac{N-1}{N}\left(\frac{c_1}{2}\frac{N+2}{N} + c_2\right).$$

This sensitivity k_1 is increasing in both c_1 and c_2. When c_1 and c_2 are increasing, market maker i's inventory cost is more sensitive to its private signal y_i. As a result, its spread s_i will also be more sensitive to the private signal y_i. The constant term k_0 is an increasing function of all three constants c_0, c_1 and c_2. Intuitively, k_0 is increasing in c_0, as a higher expected inventory cost forces

market makers to bid wider spreads. Furthermore, k_0 is also increasing in both c_1 and c_2. Note that k_0 is the submitted spread when any market maker observes the average signal $y = 0$. When both c_1 and c_2 increase, the variation of inventory cost will be larger among market makers. As a result, the marginal cost of losing the bid from marginally increasing the spread is lower, which motivates the market maker to choose a higher spread. Intuitively, when market makers are ex-post more different from each other, they are consequently willing to choose a more aggressive equilibrium strategy. The monotonicity of the equilibrium spread also implies that the winner is always the market maker with lowest signal realization, and thus the lowest inventory cost. Order-by-order auctions therefore achieve the first-best outcome in terms of efficient allocation of the retail order, as the retail order is always matched to the market maker with the lowest inventory cost.

B. Broker's routing

In this section, we consider the market equilibrium under broker's routing. In our model, we highlight the key difference between broker's routing and order-by-order auctions as market makers' different information sets when choosing spreads. Specifically, under broker's routing, market makers do not receive accurate signals about inventory cost when they compete. In practice, brokers and market makers establish long-term relationships. Market-maker performance is evaluated in the aggregate but not order-by-order, and market makers do not have a choice in when they want to accept order flow from the broker; when a broker sends, they must fulfill the order either by internalizing the order, or paying take fees to fill the order at the exchange. Focusing on this key difference, we model broker's routing by assuming that each market maker i only receives a noisy signal about y_i when submitting the spread s_i, and they are not able to adjust their offered spreads ex-post. Formally speaking, there is an additional stage, time -1, at which each market maker i receives a signal w_i. The signal w_i has the following structure. With probability p_0, $w_i = y_i$; and with probability $1 - p_0$, w_i is drawn from a uniform distribution $U\left[-\frac{1}{2}, \frac{1}{2}\right]$ which is uninformative and is independent of all other variables in the model. Each market maker i does not know whether w_i equals to y_i or not, and only understands that $w_i = y_i$ with probability p_0. Under broker's routing, all market makers submit their spreads at the end of time -1.

We still focus on symmetric equilibria in this case. In the model of broker's routing, each market maker i only observes imperfect signal w_i when they submit their spread $t_i(w_i)$. Similar to our

discussion in order-by-order auctions, we conjecture (and verify later) that there exists a linear symmetric equilibria where all market makers choose the same strategy $t(w)$, where

$$t = K_0 + K_1 w.$$

We refer readers to the appendix for more details and only present the equilibrium result.

Proposition 2. *In the model of broker's routing, there exists a linear symmetric equilibrium in which the spread submitted by market maker $i \in \{1, 2, ...N\}$ is*

$$t(w_i) = K_0 + K_1 w_i,$$

where

$$K_0 = c_0 + \frac{p_0}{4N^2}\left[\left(3 + N^2 - p_0 - Np_0\right)c_1 + 2Nc_2\right]$$

and

$$K_1 = \frac{N-1}{N}\left(c_2 p_0 + \frac{2c_1 p_0}{N} + \frac{c_1(N-2)p_0}{2N} + \frac{c_1(1-p_0)p_0}{2N}\right).$$

Proof. See Appendix. □

While we model broker's routing as a form of auction, it can also be natural to consider quantity competition in broker's routing (Kyle (1985), Baldauf et al. (2022)). Our goal here is to set up a comparable benchmark for order-by-order auctions which feature price competition, we likewise consider price competition for the broker's routing system.

We highlight the key difference between order-by-order auctions and broker's routing as the different information environment when they compete. In our model of broker's routing, if $p_0 = 1$, it becomes the model of order-by-order auctions. At the other extreme when $p_0 = 0$, market makers are homogeneously uninformed when they submit their spreads, as they have not yet observed their private signals. As a result, Bertrand competition obtains, and all market makers will earn zero expected profit in equilibrium. Therefore, the unique symmetric equilibrium spread in this case must be

$$t_i = \mathbb{E}(c_i) = \mathbb{E}\left(c_0 + c_1 \frac{1}{N}\sum_{j=1}^{N} y_j + c_2 y_i\right) = c_0$$

for all $i \in \{1, 2, ...N\}$.

Although all market makers earn a non-negative expected profit, their ex-post profit can be positive or negative, depending on the realized inventory cost. In other words, market makers will lose money on some trades. In contrast, the realized profit in order-by-order auctions must be non-negative for all market makers for all trades. Second, under broker's routing, the order will be obtained by the market maker with lowest signal w_i, who may not be the one with lowest inventory cost as w_i is just a noisy signal of y_i. As a result, welfare loss incurs due to inefficient inventory management in equilibrium. We present more detailed welfare analysis in the next subsection.

C. Welfare analysis: Order-by-order auctions vs. broker's routing

In our model, the retail order is always executed, but inventory cost and equilibrium spreads differ between order-by-order auctions and broker's routing. We denote W_M, W_I and W_{total} as the market makers' expected profit, the investor's expected profit and the total welfare, respectively:

1. The expected total profit of all market makers W_M: the expected equilibrium spread minus the incurred inventory cost;

2. The expected total profit of the retail investor W_I: the expected negative equilibrium spread;

3. The total welfare W_{total}: the expected negative incurred inventory cost, which is $W_{total} = W_M + W_I$.

Under order-by-order auctions, the market maker with the lowest signal realization executes the order in equilibrium, so the expected total profit of all market makers is

$$W_M^{OBO} = \mathbb{E}\left\{\mathbb{E}\left[k_0 + k_1 r - c_0 - c_1 \frac{1}{N}\sum_{j=1}^{N} y_j - c_2 r | \min_i y_i = r\right]\right\}.$$

The investor's expected profit is

$$W_I^{OBO} = -\mathbb{E}\left\{\mathbb{E}\left[k_0 + k_1 r | \min_i y_i = r\right]\right\},$$

and the total welfare is

$$W_{total}^{OBO} = \mathbb{E}\left\{\mathbb{E}\left[-c_0 - c_1\frac{1}{N}\sum_{j=1}^{N} y_j - c_2 r|\min_i y_i = r\right]\right\}.$$

Total welfare is the sum of the welfare of market makers and investors: $W_{total}^{OBO} = W_M^{OBO} + W_I^{OBO}$. Based on our equilibrium results, we obtain the following Lemma.

Lemma 1. *Under order-by-order auctions, the welfare outcomes of the equilibrium characterized by Proposition 1 are*

$$W_M^{OBO} = \frac{1}{N+1}\left(\frac{c_1}{N} + c_2\right),$$

$$W_I^{OBO} = -\left[c_0 + \frac{1}{N(N+1)}c_1 - \frac{N-3}{2(N+1)}c_2\right],$$

$$W_{total}^{OBO} = -\left(c_0 - \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

Proof. See Appendix. □

Order-by-order auctions implement the first best allocation, and the total welfare is

$$W_{total}^{OBO} = -\left(c_0 - \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

W_{total}^{OBO} is decreasing in the expected inventory cost c_0. Total welfare under order-by-order auctions W_{total}^{OBO} is increasing in c_2, because c_2 determines the variation of inventory cost among all market makers. When c_2 is higher, the expected lowest inventory cost will be lower, and thus the total welfare is higher. W_{total}^{OBO} is independent of c_1, because the aggregate component

$$c_1\frac{1}{N}\sum_{j=1}^{N} y_j$$

in the inventory cost always has zero mean. That is, aggregate contribution of the second component in the inventory cost is always zero, no matter how large is c_1.

Costs incurred from the parameter c_0 are borne exclusively by the investor and do not factor into market makers' welfare. An increase in the aggregate liquidity parameter c_1 leads to an improvement in market makers' welfare as each market maker's private information becomes more

relevant in the calculation of inventory costs, resulting in a more diverse bidding strategy. This, in turn, leads to market makers earning a higher information rent from the auction. Since c_1 has no impact on total welfare, when c_1 increases, investor welfare will decrease due to market makers earning higher information rents. On the other hand, an increase in c_2 has the same impact on both the investor's and market makers' welfare. Both parties benefit from an increase in c_2 as it reduces the correlation in market makers' inventory costs, leading to an overall improvement in total welfare, which is shared between the investor and market makers.

Under broker's routing, all welfare calculations are similar, except that market makers only observe noisy signals about y_i. For simplicity of exposition, we skip the intermediate steps and only present the final results.

Lemma 2. *Under broker's routing, the welfare outcomes of the equilibrium characterized by Proposition 2 are*

$$W_M^{BR} = \frac{p_0\left(2c_1 - p_0c_1 + Nc_2\right)}{N\left(1+N\right)},$$

$$W_I^{BR} = -\left[c_0 + p_0\frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)}\right],$$

$$W_{total}^{BR} = W_M^{BR} + W_I^{BR} = -\left(c_0 - p_0\frac{N-1}{N+1}\frac{c_2}{2}\right).$$

Proof. See Appendix. □

Having solved for welfare outcomes in the model of broker's routing setting and order-by-order auctions, we next compare welfare between the two systems in the following proposition.

Proposition 3. $W_{total}^{BR} < W_{total}^{OBO}$; $W_M^{BR} < W_M^{OBO}$; $W_I^{BR} < W_I^{OBO}$ *if and only if* $\frac{c_2}{c_1} > \frac{2(1-p_0)}{N(N-3)}$.

Proof. See Appendix. □

Proposition 3 is a direct result of Lemma 2. Note that the only aggregate welfare loss in this setting is from inefficient inventory management. The total welfare improvement

$$W_{total}^{OBO} - W_{total}^{BR} = (1-p_0)\frac{N-1}{N+1}\frac{c_2}{2}$$

is increasing in N and decreasing in p_0. Intuitively, when the ex-ante signal is less noisy (p_0 is higher), the order is more likely to be obtained by the market maker with the lowest inventory cost, and thus the welfare loss will be lower. The magnitude of welfare improvement also depends on the number of market makers. When there are more maker makers, the first best allocation will be more efficient as their inventory costs are not perfectly correlated. Order-by-order auctions implement the first best outcome, while the outcome of broker's routing depends on the precision of the ex-ante signal, and is less sensitive to the number of market makers. Consequently, the welfare improvement from broker's routing to order-by-order auctions is higher when there are more market makers. Lastly, the welfare improvement is also increasing in c_2, as it determines the importance of allocative efficiency gain from routing the order to the lowest-cost dealer. The broker's routing system, in contrast, is less sensitive to c_2, as it also depends on the noise from the ex-ante signal.

Market maker profit W_M is higher under order-by-order auctions, and the difference is:

$$W_M^{OBO} - W_M^{BR} = \frac{1}{N+1}\left((1-p_0)^2\frac{c_1}{N} + (1-p_0)c_2\right).$$

The difference in market-maker welfare $W_M^{OBO} - W_M^{BR}$ is increasing in c_1 and c_2. When c_1 and c_2 are higher, the private signals that market makers observe become more important in their inventory cost. Market makers are, effectively, more different ex-post. The ex-post heterogeneity generates the expected positive profit they earn under order-by-order auctions. The difference $W_M^{OBO} - W_M^{BR}$ is also increasing in $(1-p_0)$, as the precision of the noisy signal in the model of broker's routing determines the competitiveness of the market. When the signal is noisier, i.e., $(1-p_0)$ is higher, the market under broker's routing is more competitive, resulting in a lower expected equilibrium spread, and thus the difference in spreads under these two mechanisms will be larger.

The contrast between W_I^{BR} and W_I^{OBO} depends on the level of $\frac{c_1}{c_2}$, reflecting the trade-off between more efficient inventory management and higher rent earned by market makers. Since there is a common-value component in the inventory cost, and market participants' information is independent, they bid conservatively in equilibrium due to the strategic concern of the winner's curse problem. This gives market participants positive expected profit in equilibrium, which in turn hurts the investor's welfare. This effect is more severe when market participants' information

is closer to the true liquidity signal, as verified by the following result:

$$\frac{\partial^2 W_I^{BR}}{\partial c_1 \partial p_0} = -\frac{2(1-p_0)}{N(1+N)} < 0.$$

When the parameter of the common-value component c_1 increases, the investor's welfare will decrease. The above result shows that this effect is more severe when the precision p_0 is higher. Note that our order-by-order auction model is equivalent to the broker's routing model when $p_0 = 1$, this implies that when c_1 is higher, investor's welfare is more likely to be lower under order-by-order auctions, which is our prediction in Proposition 3.

A direct result from Proposition 3 is that switching to order-by-order auctions has heterogeneous impacts on stocks with different inventory cost structures. Compared to small, illiquid stocks, large, liquid stocks usually can be executed by the market makers and thus rely less on the interdealer market.[3] As a result, $\frac{c_2}{c_1}$ will be relatively larger for large liquid stocks and the smaller stock is more likely to breach the threshold $\frac{2(1-p_0)}{N(N-3)}$. Order-by-order auctions are therefore more likely to harm investor welfare in small illiquid stocks compared to large liquid stocks.

We do not directly model the endogenous entry of market makers, but our model gives implications for how market competition and liquidity provision change welfare outcomes in partial equilibrium analysis. Proposition 3 implies that when the number of market makers N is small, the investor's welfare is likely lower upon switching to order-by-order auctions. Here N measures the number of active market makers who provide liquidity. During time periods when market makers are not willing to provide liquidity (for example, due to market uncertainty or high inventory cost), our model predicts that switching to order-by-order auctions will be more likely to hurt investors. If investor protection is more important during market distress, our result highlights the unintended negative effect of order-by-order auctions during time periods when liquidity provision is limited.

D. The role of institutional traders

In the order-by-order proposal released by the SEC, the entry of institutional traders has been highlighted as a key feature of order-by-order auctions.[4] The SEC hopes that, relative to the

[3] See a microfoundation of this intuition in the appendix.

[4] As the SEC chairman Gary Gensler mentioned, "...individual investors don't necessarily get the best prices that they could get if institutional investors, like pension funds, could systematically and directly compete for their orders."

current broker's routing system, order-by-order auctions will allow institutional traders to increase competition for retail trades.[5] This hope, however, ignores the fact that institutional traders usually have superior information about asset quality compared to wholesalers (eg. Glosten and Milgrom (1985)). Allowing institutional traders to compete for retail orders may increase informational asymmetry among bidders in order-by-order auctions, and lead to a less efficient equilibrium outcome. In this section, we build a model to examine this extension and show that the entry of institutional investors brings in more adverse selection, can harm market outcomes.

To extend our model to include institutional traders, we make two (minimal) changes in the baseline model. First, apart from the N wholesalers[6] who always provide market-making service, there are $N_0 \geq 2$ institutional traders who can also provide liquidity only in order-by-order auctions. This is consistent with the market design suggested in the SEC proposal, in which institutional investors are absent in the current broker's routing system, but can be active and provide more competition in order-by-order auctions. We assume that institutional traders $i \in \{1, 2, ...N_0\}$ also receive i.i.d private signals y_i at time 0, which follows uniform distribution $U\left[-\frac{1}{2}, \frac{1}{2}\right]$. The private signal y_i plays a similar role as that for wholesalers, as discussed later in the model.

Second, we consider the following (new) inventory cost structure

$$\tilde{\zeta}_i = \tilde{c}_0 + c_1 \frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{N}} y_j + c_2 y_i, \tag{1}$$

where $\tilde{c}_0$ is a random variable that can be $c_0 - \delta_c$ or $c_0 + \delta_c$[7] with equal probabilities, and $\tilde{N}$ is the number of active market makers. In broker's routing, we only have wholesalers competing for retail orders, so $\tilde{N} = N$; and in order-by-order auctions, both wholesalers and institutional traders can compete for retail orders, so $\tilde{N} = N + N_0$ in this case. Note that $\mathbb{E}(\tilde{c}_0) = c_0$, that is, the unconditional expectation of inventory cost remains the same in this extension.

Institutional traders have an information advantage over wholesalers. Specifically, all institutional traders can observe the realization of $\tilde{c}_0$ at time 0, while wholesalers only know the distribution

Gensler (2021)

[5] A "wholesaler is often chosen by a formula that depends on past execution quality of the wholesaler, its relationship with the broker-dealer, and other factors. In addition, the bilateral nature of the wholesaler business model not only restricts contemporaneous competition among wholesalers, it also restricts opportunities for other market participants" Securities and Exchange Commission (2022).

[6] These are the market makers in our baseline model.

[7] Without loss of generality, we assume $\delta_c \geq 0$.

of $\tilde{c}_0$. This implies that when competing for retail orders, institutional traders can condition their bids on the realization of $\tilde{c}_0$, while wholesalers can only use distributional information of $\tilde{c}_0$.[8] This information asymmetry captures the nature that institutional traders are more informed of the characteristics of assets traded, market conditions or future price movement, and can change wholesalers' behavior in equilibrium due to concern about the adverse selection problem.

Let's first consider the market equilibrium in the broker's routing system. Since institutional investors are absent in broker's routing, the only difference between this extension and our baseline model is the structure of inventory cost. The additional randomness in the inventory cost (1) has no impact on market equilibrium, because all wholesalers are risk neutral and thus only care about the expectation of the $\tilde{c}_0$. Recall that $\mathbb{E}(\tilde{c}_0) = c_0$, which is the same as in the baseline model.

Proposition 4. *With inventory cost structure* (1)*, under broker's routing, the equilibrium bidding strategies and welfare outcomes are the same as characterized by Proposition 2 and Lemma 2.*

The market equilibrium is unchanged under broker's routing, thus we view it as a suitable benchmark of the (new) model with institutional traders. However, the equilibrium does change under order-by-order auctions due to the entry of institutional traders. First, consider the case when $\delta_c = 0$, when institutional traders have no informational advantage compared to wholesalers. In this case, the only effect is enhanced the competition in order-by-order auctions, which is a direct result of the increased number of bidders competing for the retail order. With Proposition 1 obtained in our baseline model, to obtain the new market equilibrium, we simply replace the number of bidders N in the baseline model with $(N + N_o)$, because institutional traders are ex-ante identical to wholesalers in this special case. The following Proposition characterizes the equilibrium strategies.

Proposition 5. *When there are N_0 institutional traders and $\delta_c = 0$, under order-by-order auctions, there exists a linear symmetric equilibrium in which the spread submitted by wholesaler or institutional trader i is*

$$\tilde{s}_i(y_i) = \tilde{k}_0 + \tilde{k}_1 y_i$$

[8] We can also interpret $\pm\delta_c$ as private information of asset quality, and keep the inventory cost structure unchanged. This will not change our model outcomes.

where

$$\tilde{k}_0 = c_0 + \frac{c_1}{4\left(N+N_0\right)}\left(N+N_o-1+\frac{2}{N+N_0}\right)+\frac{c_2}{2\left(N+N_0\right)}$$

and

$$\tilde{k}_1 = \frac{N+N_0-1}{N+N_0}\left(\frac{c_1}{2}\frac{N+N_0+2}{N+N_0}+c_2\right).$$

We consider the total welfare $\tilde{W}^{OBO}_{total}$, the investor's welfare $\tilde{W}^{OBO}_{I}$, and wholesalers' welfare $\tilde{W}^{OBO}_{W}$. Institutional traders' welfare $\tilde{W}^{OBO}_{IT}$ satisfies

$$\tilde{W}^{OBO}_{IT} = \tilde{W}^{OBO}_{total} - \tilde{W}^{OBO}_{I} - \tilde{W}^{OBO}_{W}.$$

Denote the total welfare, the investor's welfare, and wholesalers' welfare under broker's routing as $\tilde{W}^{BR}_{total}$, $\tilde{W}^{BR}_{I}$, and $\tilde{W}^{BR}_{W}$, respectively. We then compare welfare outcomes under broker's routing and order-by-order auctions in this extension.

Proposition 6. *When there are N_0 institutional traders and $\delta_c = 0$, we have the following results on welfare comparison:*

1. $\tilde{W}^{BR}_{total} < \tilde{W}^{OBO}_{total}$;
2. $\tilde{W}^{BR}_{W} < \tilde{W}^{OBO}_{W}$ *if and only if* $\frac{N(N+1)}{(N+N_0)(N+N_0+1)} > p_0$ *and* $\frac{c_2}{c_1} > -\frac{1}{N+N_0}\frac{\frac{N(N+1)}{(N+N_0)(N+N_0+1)} - p_0\frac{(N+N_0)(2-p_0)}{N}}{\frac{N(N+1)}{(N+N_0)(N+N_0+1)} - p_0}$;
3. $\tilde{W}^{BR}_{I} < \tilde{W}^{OBO}_{I}$ *if and only if* $\frac{c_2}{c_1} > \frac{\frac{1}{(N+N_0)(1+N+N_0)} - \frac{p_0(2-p_0)}{N(N+1)}}{\frac{N+N_0-3}{2(N+N_0+1)} - \frac{p_0(N-3)}{2(N+1)}}$.

When institutional traders provide liquidity in order-by-order auctions but have no informational advantage, the total welfare unambiguously improves when switching from broker's routing to order-by-order auctions. Since the order is always obtained by the market maker with the lowest ex-post inventory cost under order-by-order auctions and their inventory costs are not perfectly correlated, having institutional traders in order-by-order auctions always makes the order allocation more efficient. The effect on investor's welfare is ambiguous, which is higher under order-by-order auctions if and only if $\frac{c_2}{c_1}$ is greater than a threshold

$$\frac{\frac{1}{(N+N_0)(1+N+N_0)} - \frac{p_0(2-p_0)}{N(N+1)}}{\frac{N+N_0-3}{2(N+N_0+1)} - \frac{p_0(N-3)}{2(N+1)}},$$

qualitatively similar to the Proposition 3 in the baseline model. If

$$\frac{1}{(N+N_0)(1+N+N_0)} - \frac{p_0(2-p_0)}{N(N+1)} < 0, \tag{2}$$

the threshold is always negative, and the investor's welfare always improve under order-by-order auctions, irrespective of the level of $\frac{c_2}{c_1}$. Condition (2) concerns the number of new institutional traders providing liquidity under order-by-order auctions. Under the joint assumption that institutional traders have no informational advantage when they compete for retail orders (i.e., when $\delta_c = 0$) and that order-by-order auctions can attract sufficiently many institutional traders, investors will unambiguously benefit from switching to order-by-order auctions, as the benefit of efficient inventory management will dominate any decreases in competition. This is precisely the intuition motivating the SEC's proposal on order-by-order auctions, and our above results highlight the underlying assumptions required for it to hold.

After switching to order-by-order auctions, the wholesalers' welfare is increasing if and only if two conditions are satisfied. First, the number of new institutional investors N_0 has to be low enough, i.e.,

$$\frac{N(N+1)}{(N+N_0)(N+N_0+1)} > p_0.$$

Unconditionally, all wholesalers and institutional traders can obtain the order with equal probabilities. When there are sufficiently many institutional investors, the wholesalers' welfare mechanically decreases due to the competition. When there are sufficiently many new institutional traders in order-by-order auctions, the wholesalers will surely be worse off.

Second, $\frac{c_2}{c_1}$ must exceed the threshold:

$$\frac{c_2}{c_1} > \frac{-\frac{1}{N+N_0}\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)}{\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}}. \tag{3}$$

Note that wholesalers' welfare unambiguously improves from broker's routing to order-by-order auctions in our baseline model. However, with the entry of institutional traders, the wholesalers' welfare increases only when

$$\frac{c_2}{c_1}$$

is sufficiently high, because the entry of new institutional traders also decreases the total welfare of all market makers (wholesalers and institutional traders). When $\frac{c_2}{c_1}$ is sufficiently high, the market makers' inventory cost will be more heterogeneous ex-post, creating more information rent for them. So the wholesalers' welfare is higher under order-by-order auctions only when $\frac{c_2}{c_1}$ is sufficiently high.

We next consider the case when institutional investors have at least some informational advantage, that is, when $\delta_c > 0$. We still focus on symmetric linear strategies where all institutional traders choose the same linear strategy and all wholesalers choose the same linear strategy. When $\delta_c > 0$, institutional traders can condition their bids on the realization of $\tilde{c}_0$. Intuitively, when observing $\tilde{c}_0 = c_0 - \delta_c$, institutional traders will submit lower bids, and when $\tilde{c}_0 = c_0 + \delta_c$, they will submit higher bids. In contrast, wholesalers cannot condition their spreads on the realizations of $\tilde{c}_0$, but just the distributional information c_0. Wholesalers are more likely to win auctions when $\tilde{c}_0 = c_0 + \delta_c$ than when $\tilde{c}_0 = c_0 - \delta_c$, as institutional traders will bid more aggressively in the latter case. This leads to adverse selection for wholesalers, as they are more likely to win auctions when $\tilde{c}_0 > E(\tilde{c}_0)$. A winner's curse argument implies that wholesalers will submit more conservative bids in equilibrium. When δ_c is sufficiently large, the winner's curse concern becomes so severe, such that all wholesalers will be completely out of competition for high-quality (low-cost) stocks, and can only obtain the retail order when $\tilde{c}_0 = c_0 + \delta_c$. Consequently, when $\tilde{c}_0 = c_0 - \delta_c$, institutional traders will face no competition from wholesalers, which can reduce retail investor welfare. The following proposition formalizes this intuition:

Proposition 7. *Let $\tilde{s}^- (y; \delta_c)$ and $\tilde{s}^+ (y; \delta_c)$ be two bidding strategies, where*

$$\tilde{s}^- (y; \delta_c) = \tilde{k}_0^- (\delta_c) + \tilde{k}_1^- (\delta_c) y_i$$

with

$$\tilde{k}_0^- (\delta_c) = c_0 - \delta_c + \frac{c_1}{4N_0} \left(N_o - 1 + \frac{2}{N_0} \right) + \frac{c_2}{2N_0}$$

$$\tilde{k}_1^- (\delta_c) = \frac{N_0 - 1}{N_0} \left(\frac{c_1}{2} \frac{N_0 + 2}{N_0} + c_2 \right),$$

and

$$\tilde{b}^+ (y; \delta_c) = \tilde{k}_0^+ (\delta_c) + \tilde{k}_1^+ (\delta_c) y$$

with

$$\tilde{k}_0^+(\delta_c) = c_0 + \delta_c + \frac{c_1}{4(N+N_0)}\left(N+N_0-1+\frac{2}{N+N_0}\right) + \frac{c_2}{2(N+N_0)}$$

$$\tilde{k}_1^+(\delta_c) = \frac{N+N_0-1}{N+N_0}\left(\frac{c_1}{2}\frac{N+N_0+2}{N+N_0}+c_2\right).$$

When there are N_0 institutional traders, there exists a threshold $\underline{\delta} > 0$, such that when $\delta_c > \underline{\delta}$, there exists an equilibrium of order-by-order auctions in which

1. *the wholesalers always choose bidding strategy $\tilde{s}^+(y;\delta_c)$;*
2. *institutional traders choose bidding strategy $\tilde{s}^+(y;\delta_c)$ when observing $c_0+\delta_c$ and $\tilde{s}^-(y;\delta_c)$ when observing $c_0-\delta_c$.*

The threshold $\underline{\delta}$ satisfies the following condition

$$\tilde{k}_0^- + \tilde{k}_1^-\frac{1}{2} < \tilde{k}_0^+ - \tilde{k}_1^+\frac{1}{2}.$$

This implies that when the true state is $\tilde{c}_0 = c_0 - \delta_c$, the highest possible spread offered by institutional traders is still lower than the lowest possible spread offered by wholesalers, and thus wholesalers will never obtain the order in this case, irrespective of their signal realizations. When the true state is $\tilde{c}_0 = c_0+\delta_c$, wholesalers and institutional traders will choose the symmetric bidding strategy $\tilde{b}^+(y;\delta_c)$, and thus all players will obtain the order with equal probabilities in this case.

If we interpret the random variable $\tilde{c}_0$ as the heterogeneous quality of stocks, then in equilibrium, institutional traders compete effectively only for retail orders of high-quality stocks, while all market makers compete for orders of low-quality stocks. The market for low-quality stocks becomes more competitive due to an increase in the number of bidders, while the market for high-quality stocks may become less competitive as institutional traders are the only effective bidders. The presence of adverse selection can weaken competition and potentially harm total welfare, as our following proposition illustrates.

Proposition 8. *When there are N_0 institutional traders and $\delta_c > \underline{\delta}$, we have the following results on welfare comparison:*

1. *$\tilde{W}_{total}^{BR} < \tilde{W}_{total}^{OBO}$ if and only if $N_0 > \underline{N}_0$, where $\underline{N}_0$ is a constant solved in appendix by* (B6)*;*

2. $\tilde{W}_W^{BR} < \tilde{W}_W^{OBO}$ *if and only if* $p_0 < \frac{1}{2}\frac{N}{N+N_0}\frac{1+N}{N+N_0+1}$ *and* $\frac{c_2}{c_1} > \frac{-\frac{1}{N+N_0}\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)}{\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}}$;

3. $\tilde{W}_I^{BR} < \tilde{W}_I^{OBO}$ *if and only if* $p_0 < \frac{1-\frac{2}{N+N_0+1}-\frac{2}{N_0+1}}{1-\frac{4}{N+1}}$ *and* $\frac{c_2}{c_1} > \frac{\frac{1}{(N+N_0)(N+N_0+1)}+\frac{1}{N_0(N_0+1)}-\frac{2p_0(2-p_0)}{N(1+N)}}{1-p_0+\frac{4p_0}{N+1}-\frac{2}{N+N_0+1}-\frac{2}{N_0+1}}$.

The comparison of total welfare between broker's routing and order-by-order auctions is complicated by the presence of adverse selection. The transition from broker's routing to order-by-order auctions results in an improvement in total welfare only when there is a sufficient number of institutional traders providing liquidity. However, the absence of wholesalers can result in a decrease in market competitiveness for high quality stocks, leading to an inefficient outcome for their trading. While low quality stocks may experience an increase in market competitiveness, this gain may not be enough to offset the welfare loss from the trading of high quality stocks.

In accordance with Proposition 6, the welfare effects on both investors and wholesalers are similar. Both parties are likely to benefit from stocks with a high $\frac{c_2}{c_1}$. As previously noted in our baseline model (and appendix), stocks with a high $\frac{c_2}{c_1}$ tend to be large and highly liquid, leading to welfare losses for small, illiquid stocks.[9]

E. Heterogeneous stocks and cross-subsidization

In our baseline model, we consider a unit order from a single stock, and the equilibrium and welfare outcomes depend on parameters (c_0, c_1, c_2). In this section, we extend our baseline model to heterogeneous stocks with different characteristics (c_0, c_1, c_2). For order-by-order auctions, this extension is straightforward, as the stock characteristics (c_0, c_1, c_2) is publicly observable when market makers compete. Then the market equilibrium (spread, allocation, and welfare outcomes) can still be captured by our baseline model. The extension, however, is less straightforward for broker's routing. This is because broker's routing features the long-term relationship between brokers and market makers, and thus the competition among market makers happens before the order actually arrives and the order characteristics are observed. As a result, market outcomes among heterogeneous stocks under broker's routing will be less differentiated compared to that under order-by-order auctions. Our model predicts that compared to order-by-order auctions, there is less variation in equilibrium spreads among stocks under broker's routing. Based on

[9]This is also in line with the concerns expressed by practitioners, who generally believe that the transition to order-by-order auctions may negatively impact small and illiquid stocks.

this observation, we can also highlight a cross-subsidization effect: under broker's routing, the equilibrium spreads of high-cost stocks are relatively low (compared to that under order-by-order auctions), while the equilibrium spreads of low-cost stocks are relatively high. This cross-subsidization effect implies that switching from broker's routing to order-by-order auctions not only changes the retail investors' total welfare, but also changes the welfare distribution when retail investors have different portfolio holdings.

To capture this idea, we consider a pool of orders characterized by a joint cumulative distribution $G(c_0, c_1, c_2)$, and the realization of (c_0, c_1, c_2) is independent of all other variables in the model. For simplicity, we assume that G has full support on $(0, \infty) \times (0, \infty) \times (0, \infty)$, and is continuously differentiable everywhere. We consider a model with the following timeline:

1. At time -1, the cumulative distribution function $G(c_0, c_1, c_2)$ becomes public information, and each market maker i observes his private noisy signal w_i;

2. At time 0, an order with characteristics (c_0, c_1, c_2) is drawn from distribution G, and each market maker i observes his private signal y_i. The broker then sends the order (c_0, c_1, c_2) to one market maker which is determined by the allocation mechanism;

3. At time 1, all random variables are realized and all market participants collect their payoffs.

As we discussed in the baseline model, under broker's routing, market makers compete and submit their spreads at time -1, while under order-by-order competition, they submit their spreads at time 0. Let's first introduce the following variables

$$\bar{c}_0 = \iiint c_0 dG(c_0, c_1, c_2) = \mathbb{E}(c_0),$$

$$\bar{c}_1 = \iiint c_1 dG(c_0, c_1, c_2) = \mathbb{E}(c_1),$$

$$\bar{c}_2 = \iiint c_2 dG(c_0, c_1, c_2) = \mathbb{E}(c_2).$$

Under order-by-order auctions, since order characteristics (c_0, c_1, c_2) are public, the equilibrium and welfare outcomes are the same as characterized by Lemma 1 and Lemma 2 in our baseline model.

Under broker's routing, since only distributional information G is available when market makers compete at time -1, the equilibrium strategy will only depend on the distributional information G but not the specific order characteristics (c_0, c_1, c_2). The new equilibrium of broker's routing is characterized by the following Proposition.

Proposition 9. *In the extension of heterogeneous stocks, under broker's routing, there exists an equilibrium in which every market maker who observes signal w chooses to submit spread*

$$\bar{T}(w) = \bar{K}_0 + \bar{K}_1 w$$

where

$$\bar{K}_0 = \bar{c}_0 + \frac{p_0}{4N^2}\left[\left(3 + N^2 - p_0 - Np_0\right)\bar{c}_1 + 2N\bar{c}_2\right]$$

and

$$\bar{K}_1 = \frac{N-1}{N}\left(\bar{c}_2 p_0 + \frac{2\bar{c}_1 p_0}{N} + \frac{\bar{c}_1 (N-2) p_0}{2N} + \frac{\bar{c}_1 (1-p_0) p_0}{2N}\right).$$

Since all market makers are risk neutral and the equilibrium is linear in the baseline model, we still obtain a linear equilibrium in this extension. Consider K_0 and K_1 in the baseline model as functions of (c_0, c_1, c_2), the equilibrium strategy in this extension satisfies

$$\bar{T}(w) = \mathbb{E}(t(w)) = \mathbb{E}(K_0 + K_1 w) = \bar{K}_0 + \bar{K}_1 w.$$

Then market makers choose an average bidding strategy in this extension. Note that both K_0 and K_1 are increasing functions of c_0, c_1 and c_2, this result implies that, compared to our baseline model results, the equilibrium spread in this extension is relatively low for stocks with high inventory cost characteristics, and high for stocks with low inventory cost characteristics.

The welfare impacts impacts are also heterogeneous. To be specific, we consider the welfare outcomes for any specific order with characteristics (c_0, c_1, c_2). The following Lemma summarizes our results.

Lemma 3. *In the equilibrium characterized by Proposition 9, the investor's welfare $\bar{W}_I^{BR}$, the total*

welfare $\bar{W}^{BR}_{total}$ *and market makers' welfare* $\bar{W}^{BR}_{M}$ *are*

$$\bar{W}^{BR}_{heter,I} = -\left[\bar{c}_0 + p_0 \frac{2\left(2-p_0\right)\bar{c}_1 - \left(N-3\right)N\bar{c}_2}{2N\left(1+N\right)}\right],$$

$$\bar{W}^{BR}_{heter,total} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right),$$

$$\begin{aligned}\bar{W}^{BR}_{heter,M} &= \bar{W}^{BR}_{heter,total} - \bar{W}^{BR}_{heter,I} \\ &= \left(\bar{c}_0 - c_0\right) + \frac{p_0}{2\left(N+1\right)}\left[\left(N-1\right)c_2 - \left(N-3\right)\bar{c}_2\right] + p_0\frac{\left(2-p_0\right)\bar{c}_1}{N\left(1+N\right)}.\end{aligned}$$

The total welfare in Lemma 3 is the same as that in the baseline model. Note that in our model, the total welfare is only determined by inventory cost but not the equilibrium spread, as the spread is just a transfer between market makers and the investor. Since the order is always obtained by the market maker with the lowest signal w_i, and introducing heterogeneity in stocks does not change allocative efficiency, we conclude that the total welfare is the same as that in the baseline model for any order (c_0, c_1, c_2) in this extension. However, the equilibrium spread does change. Specifically, now the investor's welfare (which is the negative expected equilibrium spread) becomes

$$-\left[\bar{c}_0 + p_0 \frac{2\left(2-p_0\right)\bar{c}_1 - \left(N-3\right)N\bar{c}_2}{2N\left(1+N\right)}\right]$$

which only depends on the average levels $(\bar{c}_0, \bar{c}_1, \bar{c}_2)$ but not order characteristics (c_0, c_1, c_2). Note that under order-by-order auctions, the investor's welfare is

$$-\left[c_0 + \frac{1}{N(N+1)}c_1 - \frac{N-3}{2(N+1)}c_2\right]$$

which depends on order characteristics (c_0, c_1, c_2). Then investors will be worse off after switching to order-by-order auctions if c_0 is high, c_1 is high, or c_2 is low. This highlights our cross-subsidization effect under broker's routing that market makers charge low equilibrium spreads for high-cost stocks and high equilibrium spreads for low-cost stocks. This cross-subsidization effect also implies that switching from broker's routing to order-by-order auctions may have unintended effects on retail investors' welfare distribution. For example, investors who mainly trade small, illiquid stocks with

high average inventory cost c_0 will be worse off after switching to order-by-order auctions, while those who trade large, liquid stocks with low average inventory cost c_0 will be better off.

The market maker's welfare is

$$(\bar{c}_0 - c_0) + \frac{p_0}{2(N+1)}\left[(N-1)c_2 - (N-3)\bar{c}_2\right] + p_0\frac{(2-p_0)\bar{c}_1}{N(1+N)}$$

which depends on the difference between order characteristics (c_0, c_1, c_2) and the average levels $(\bar{c}_0, \bar{c}_1, \bar{c}_2)$. Under order-by-order auctions, the market maker welfare is

$$\frac{1}{N+1}\left(\frac{c_1}{N} + c_2\right)$$

which is always positive. However, under broker's routing, marker makers make more profit from stocks with relatively low inventory cost, and incur loss from stocks with high average inventory cost, and the welfare (or net profit) from executing a specific order may be negative. This is consistent with our observation that under the current broker's routing system, market makers sometimes lose by providing liquidity for small, illiquid stocks but they can make a profit from executing large liquid stocks. On average, they can make positive expected profit from market making. Our result implies that, after switching to order-by-order auctions, market makers will only submit spreads that are high enough such that they can earn a positive expected profit on every individual order.

IV. Retail Liquidity Programs

Several exchanges have developed retail liquidity programs (RLP). These programs enable market makers to enter hidden limit orders which improve on the NBBO but are only accessible to retail orders. These hidden limit orders can be priced in any one-tenth of a cent increment, with the exception of the IEX RLP which only allows pricing at the mid-quote. While the orders are hidden, if there is more than one round lot of interest at a price which improves the NBBO, the exchange will disseminate an indicative flag highlighting that there is a resting limit order, but it will not indicate the price or size of the order.

Retail Liquidity Programs offer the closest existing analogue to the contemplated order-by-

order competition system for retail orders.[10] If multiple market makers place limit orders, for example, the market maker with the best priced limit order will win any incoming retail market order. In approving RLPs, the SEC itself has frequently highlighted the same objectives that it has for proposing order-by-order competition, namely to increase the number of market participants interacting with retail orders. These RLP limit orders are only accessible to incoming market orders from retail investors, preserving the segmentation of retail investors, but also having entirely voluntary participation: market makers may chose to stop bidding for incoming orders at any time. We analyze RLPs as a way to gain insight into how the order-by-order system would function, and identify similarities between our model and the current utilization of RLP programs.

A. Program Details

NYSE was the first to operate a RLP, on August 1, 2012.[11] The NYSE RLP was initially approved as a pilot and given several temporary pilot extensions until permanent approval on February 15, 2019. Any NYSE member can submit a Retail Price Improvement Order (RPI). An RPI order can be submitted in \$0.001 increments, and must improve the best bid or offer on the NYSE or NYSE Arca book by at least \$0.001. The size and exact price of resting RPI orders are non-displayed, but the orders do trigger indicative messages on the SIP and NYSE proprietary data feeds indicating whether there is any RPI interest at the ask, any RPI interest at the bid, or any RPI interest at both. Incoming marketable retail orders can trade against resting RPI orders. Incoming retail orders will first trade against the best-priced orders; if there is a non-displayed order which is not RPI at the mid-quote, the retail order would trade against the mid-quote interest before trading against any RPI orders priced between the mid-point and near side. Retail marketable orders can be set to only trade against RPI and non-displayed orders, or to trade against any RPI and non-displayed orders and then subsequently against the displayed best quotes up to the limit price.

[10] Bishop (2022) notes: "Exchanges already have ways for retail orders to be identified and treated specially by market makers, called retail liquidity programs (RLPs). The details differ across exchanges, but they typically allow market participants (including market makers and institutional investors) to submit orders that will interact solely or distinctly with retail-identified orders. Such orders operate on the continuous books of the exchanges, rather than executing via auctions. It seems that such existing mechanisms can deliver a similar benefit to retail investors through order-by-order competition among market makers and institutional investors."

[11] The introduction of the data field for the RLP led to the \$400 million trading glitch at Knight Capital Group on the first day that the new data field was active.

The NYSE Retail Liquidity Program charges no trading fee to qualifying retail market orders. The NYSE RLP program also pays $0.0003 credit to a Retail Liquidity Provider whenever their RPI limit order fills a retail market order. To qualify as a Retail Liquidity Provider on the NYSE, a firm must maintain a resting RPI order which improves the best bid or offer for at least 5% of the trading day.

This 5% rule distinguishes the NYSE Retail Liquidity Program from those offered by NASDAQ and BATS, with both competing programs being developed shortly after the NYSE program. The BATS program was approved as a pilot on November 27, 2012, while the NASDAQ program was approved as a pilot on February 15, 2013. Both programs have no requirement to provide liquidity for a certain percentage of the trading day, and are therefore potentially more accessible to non-market-making firms. In approving the NASDAQ RLP, the SEC notes that "the Program might also create a desirable opportunity for institutional investors to interact with retail order flow that they are not able to reach currently. Today, institutional investors often do not have the chance to interact with marketable retail orders that are executed pursuant to internalization arrangements. Thus, by submitting RPI Orders, institutional investors may be able to reduce their possible adverse selection costs by interacting with retail order flow" SEC (2013). The SEC identifies the same desirable feature, that of more potential counter parties for retail trades, that are highlighted in a potential move to order-by-order competition.

The Investors Exchange (IEX) offers a retail liquidity program whereby retail liquidity providers can enter hidden mid-point peg limit orders which are only available to retail market orders. All mid-point peg orders enter the same time priority queue, whether or not they are only available to retail investors, and both have queue priority over the IEX D-limit order, which is the discretionary limit order which takes advantage of the IEX speed bump to reprice when it detects a crumbling quote. The IEX RLP only takes mid-point orders, and disseminates a RLP indicative flag when there is at least one round lot of RLP interest. All eligible retail orders have no trading fees, either for the retail broker or the retail liquidity provider.

The IEX RLP is the most recent program, first offering the RLP trading functionality on October 1, 2019. IEX initially had no RLP indicators, but added indicators on October 13, 2021. Unlike other retail liquidity programs, the IEX program only allows mid-quote prices. Therefore, while the size available is hidden, an advertised RLP indicator from IEX confirms that at least 100

shares are available at the specific price of the mid-quote. To offer RLP indicators, the program required an approved exemption from SEC Rule 242.602, as the RPI would indicate a specific price and a minimum quantity of shares, but would not be accessible to non-retail marketable orders.

The Members Exchange MEMX applied to create an RLP program, but was denied by the SEC on February 14, 2022. The MEMX proposal differed from previous proposals in the determination of price-time priority. Under the MEMX proposal, incoming retail market orders first interact with hidden RPI orders before interacting with hidden non-RPI orders, even if the hidden non-RPI are at the same price level and have time priority. MEMX argued that because hidden RPI orders do contribute to the dissemination of the RPI interest indicator, they should have priority over hidden non-RPI orders at each price level, analogous to standard practice of non-hidden orders having priority over hidden orders at each price level. The SEC disagreed, and ruled that the change in priority would violate Section (6)(b)(5) and Section 11A of the Exchange Act.[12]

B. Data and Summary Statistics

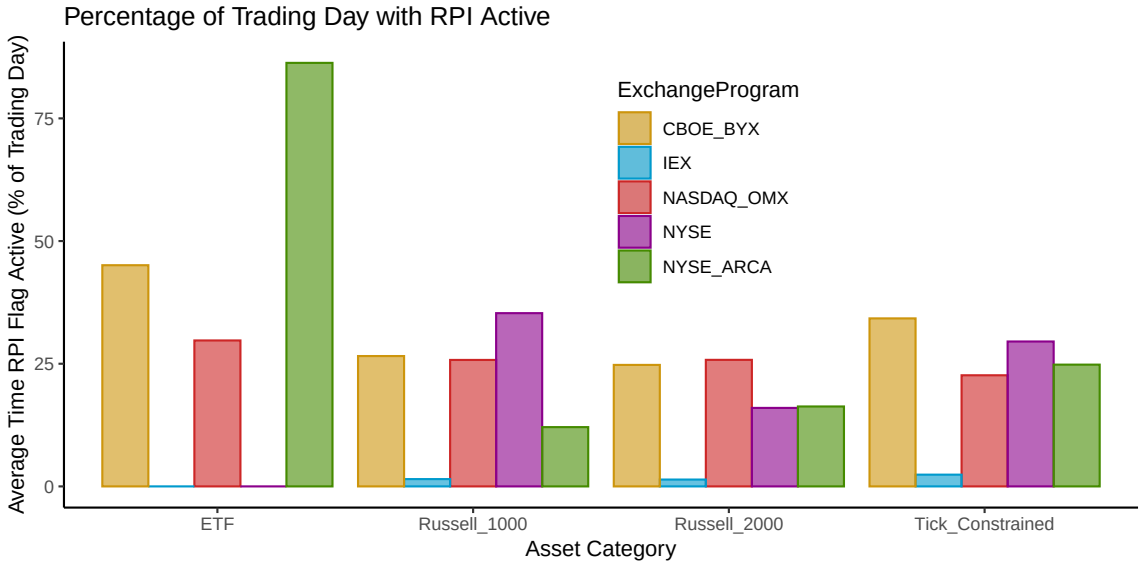
We obtain NYSE TAQ (Trade and Quote) data from January 1, 2019 to May 30, 2022. We examine all securities in the Russell 3000 index, as well as the 100 most frequently traded ETFs, provided these securities are priced above $1 per share. To exclude fractional shares, as documented by Bartlett, McCrary, and O'Hara (2022), we exclude any orders for exactly 1 share, as these may be orders for a fractional share which rounds up to 1 share.

Retail Liquidity Program (RLP) Indicators are distributed through the SIP, and are available in TAQ Data. As indicators may be disseminated even when an exchange's visible posted best bid or offer (BBO) is not at the official NBBO, we obtain retail indicator flags from the TAQ Quotes file. For each trade occurring at an exchange with an RLP, we check whether the trade occurred with an active RPI Flag by matching the RLP quotes for that exchange using the participant timestamp. We also construct an indicator for whether any RLP from the five different programs is active at any point in time, and match this to both on-exchange and off-exchange trades using the participant timestamps.

[12]Ironically, in the Order-by-Order proposal from the SEC, auctions would be required to give auction responses higher priority than hidden limit orders (Securities and Exchange Commission (2022), Proposed Rule 615 Section IV C.5.) In other words, MEMX's RLP was denied because it proposed giving resting RPI orders priority over hidden resting limit orders, but auctions would be *required* to give auction responses priority over hidden resting limit orders.

Retail Liquidity Programs have indicative interest for a large portion of the trading day. Figure 1 plots the percentage of time, by asset, that there is at least one-sided RPI interest. ETFs often have resting RPI orders for 50 to 75% of the trading day. For stocks of the Russell 1000, NYSE, CBOE, and NASDAQ have resting RPI orders for over 20% of the day. For stocks of the Russell 2000, both CBOE and NASDAQ have resting RPI orders for over 20% of the day. While some of the differences in RPI shares across assets may come from the different rules, as outlined in section V.A, there is also a considerable listing-exchange advantage. NYSE Arca's retail liquidity program, for example, has RPI interest for less than 20% of the trading day for Russell 1000 or Russell 2000 stocks, but has RPI interest for over 75% of the trading day for ETFs in our sample.

Figure 1. Time Share of Retail Liquidity Programs. We plot the average time that an RPI indicator is active, measured as percentage of time active out of the total trading day. Our sample can be divided into three groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the day for at least one-third of the days of our sample.



While the percentage of the trading day with RPI interest is considerable, the volumes executed through RPI programs are diminutive. Figure 2 depicts the trading volume split of trades when the exchange's RPI Flag is active, and the trading volume split when it is not. Sub-penny executions are less than 1% of total trading volume at exchanges, even when the RPI Flag is active. On-exchange mid-quote trading volume is considerably higher when the RPI Flag is active, but still represents less than 5% of total trading volume. Furthermore, this mid-quote volume is a mixture

of both retail interest, including from the IEX RLP which only allows retail RPI orders to be priced at mid-quote, and non-RLP program hidden mid-quote liquidity. The vast majority of mid-quote and sub-penny trading occurs off-exchange. When RPI Flags are active, a larger share of off-exchange volume occurs at sub-penny or mid-quote prices. Table I presents the exact total volumes in our sample executed when RPI programs are active, and when they are not. Note that exchange sub-penny volume when there is no RPI Flag is small but non-zero. This sub-penny volume when there is no RPI Flag can arise from hidden RLP liquidity of less than one round lot, as the RPI Flag is only disseminated when there is at least one round lot of interest. Another possible explanation for this discrepancy is inaccuracy in the timestamp-based matching of the sort described by Schwenk-Nebbe (2021), who show that the exchange processing and dissemination of quotes is typically several microseconds faster than that of trades.

There is a considerable discrepancy between the share of time that retail liquidity programs have RPL flags active, and the share of trading volume which executes in RLP. Figure 3 highlights that RPI interest is much lower in the morning, and increases throughout the day for most RLP programs. Across each time interval, the IEX RLP is active for a notably smaller percentage of time relative to any competing RLPs, as the IEX RLP requires orders to be placed at mid-quote, while competitor programs only require a minimum of 10 mils of improvement relative to the NBBO. The RLP flags also display no indication of the size available, with the flag only indicating whether there is at least one round lot.

The total volume share of Retail Liquidity Programs is stable during our sample period. As Panel A of Figure 4 depicts, on-exchange sub-penny retail trades are consistently less than 0.2% of total volume for the Russell 1000 and Russell 2000 stocks in our sample. The volume share of ETFs and tick-constrained stocks is slightly higher, at around 0.2% to 0.5% of total trading volume. We define a stock as tick-constrained if it has a one penny bid-ask spread for at least 50% of the trading day for at least one-third of the trading days in our sample. For these stocks, competition for a marketable order is potentially larger due to the tick constraint, with increased interest in providing liquidity in an RLP. In Panel B of Figure 4, we plot the volume of any exchange sub-penny or mid-quote executions while the RPI Flag is active. While this will include some non-retail hidden liquidity, it also captures retail interest at mid-quote, which is crucial as the IEX RLP only allows pricing retail price improvement at mid-quote. For ETFs and tick-constrained stocks,

Table I: Summary Volumes By Each Price Increment. This table presents summary total trading volume (in billions of dollars) in our sample for each sub-penny category of trade: at-quote, mid-quote, and sub-penny. Panel A of Figure presents volume for exchange trades. We define an RPI Flag as active if there is contemporaneous RLP interest at the exchange where trade occurs. Panel B presents the volume for off-exchange trades. Note that Panel B is off-exchange trades only, and we define the RPI Flag as active if there is contemporaneous RLP interest at any exchange with an RLP program.
Our sample can be divided into three asset groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) asset group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the the day for at least one-third of the days of our sample.

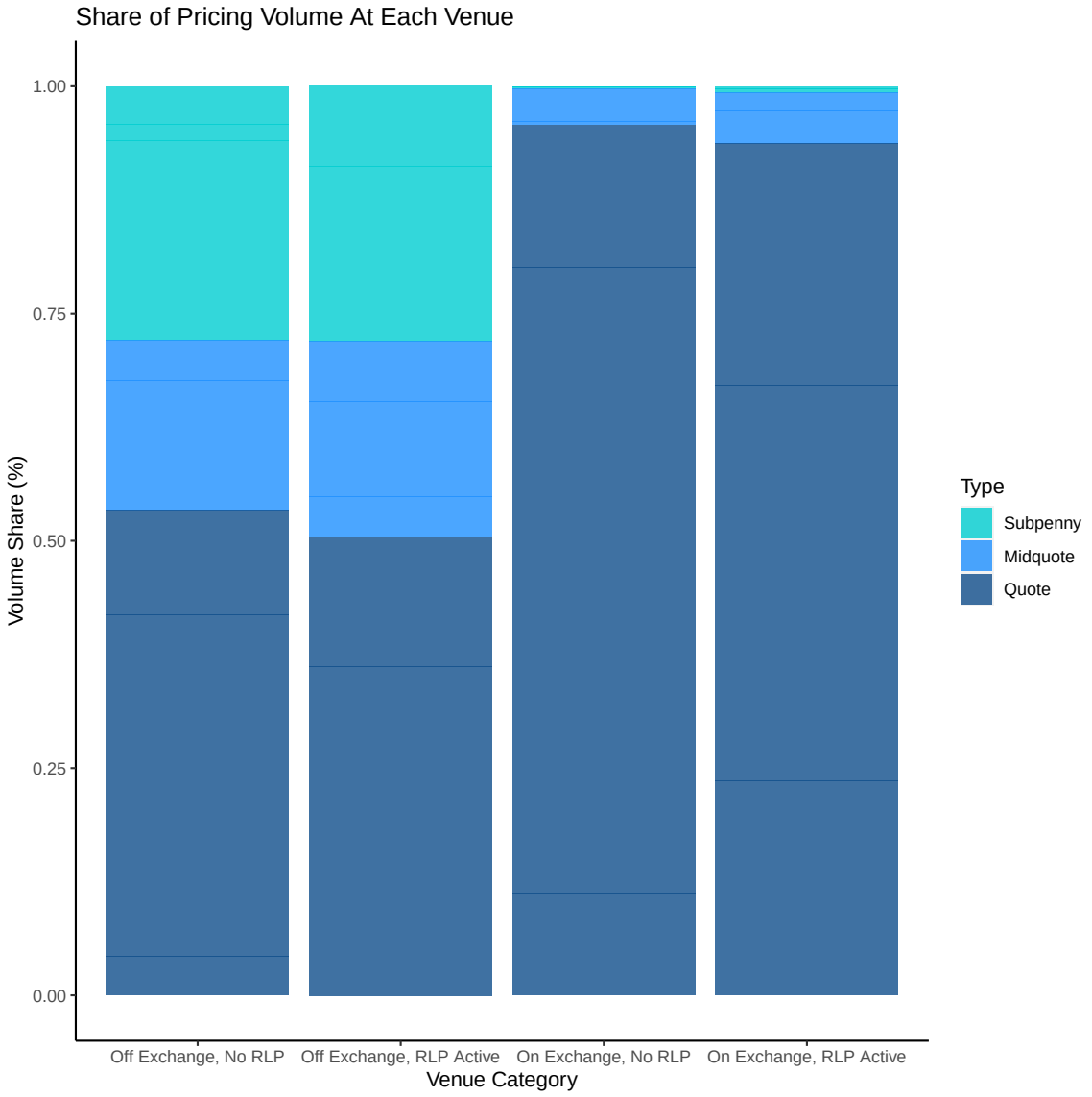
Panel A: Exchange Trades

	RLP	Volume			Percent		
Asset Class	Flag	Mid-quote	At Quote	Sub-penny	Mid-quote	At Quote	Sub-penny
ETF	Active	2259	35868	414	3.2	51	0.6
ETF	None	317	9160	55	0.5	13	0.1
Russell_1000	Active	3162	60890	235	1.7	32	0.1
Russell_1000	None	1933	50111	99	1.0	27	0.1
Russell_2000	Active	283	5376	13	1.5	28	0.1
Russell_2000	None	306	6012	2.6	1.6	31	0.01
TickConstrained	Active	2767	40510	400	3.3	48	0.5
TickConstrained	None	676	12773	67	0.8	15	0.1

Panel B: Off-Exchange Trades

	RLP	Volume			Percent		
Asset Class	Flag	Mid-quote	At Quote	Sub-penny	Mid-quote	At Quote	Sub-penny
ETF	Active	3626	9070	5751	5.2	13	8.2
ETF	None	537	1909	1134	0.8	2.7	1.6
Russell_1000	Active	8638	20795	10127	4.6	11	5.4
Russell_1000	None	5895	16953	8756	3.1	9.0	4.7
Russell_2000	Active	743	2161	822	3.8	11	4.2
Russell_2000	None	723	2107	833	3.7	11	4.3
TickConstrained	Active	4740	9643	6456	5.6	12	7.7
TickConstrained	None	1257	3079	1869	1.5	3.7	2.2

Figure 2. Volume Share of Venues. We plot the percentage of volume which executes either at the quote, at the mid-quote, or at a sub-penny price for both on-exchange and off-exchange venues. On both types of venues, a higher percentage of volume occurs at the quote when there is no RPI Flag active, and a higher share of volume executes at sub-penny and mid-quote prices when the RPI Flag is active.



exchange sub-penny and mid-quote volume when the RPI Flag is active is around 0.5% to 1.0% of trading volume. While this is a small share of total trading volume, it represents a much larger fraction of retail-only trading volume.

Figure 3. Intra-day Time Share. We plot the average share of time that the RPI Flag is active throughout the trading day on January 3, 2022. For each exchange, we divide the trading day into 30-minute intervals and calculate the average across stocks of the percentage of time for which the RPI Flag is active. One-sided liquidity is the percentage of time for which there is a quote on either the bid, the ask, or both, and therefore includes the time for which there is two-sided liquidity (i.e., a flag indicating RPI interest on both the bid and ask at the same time).

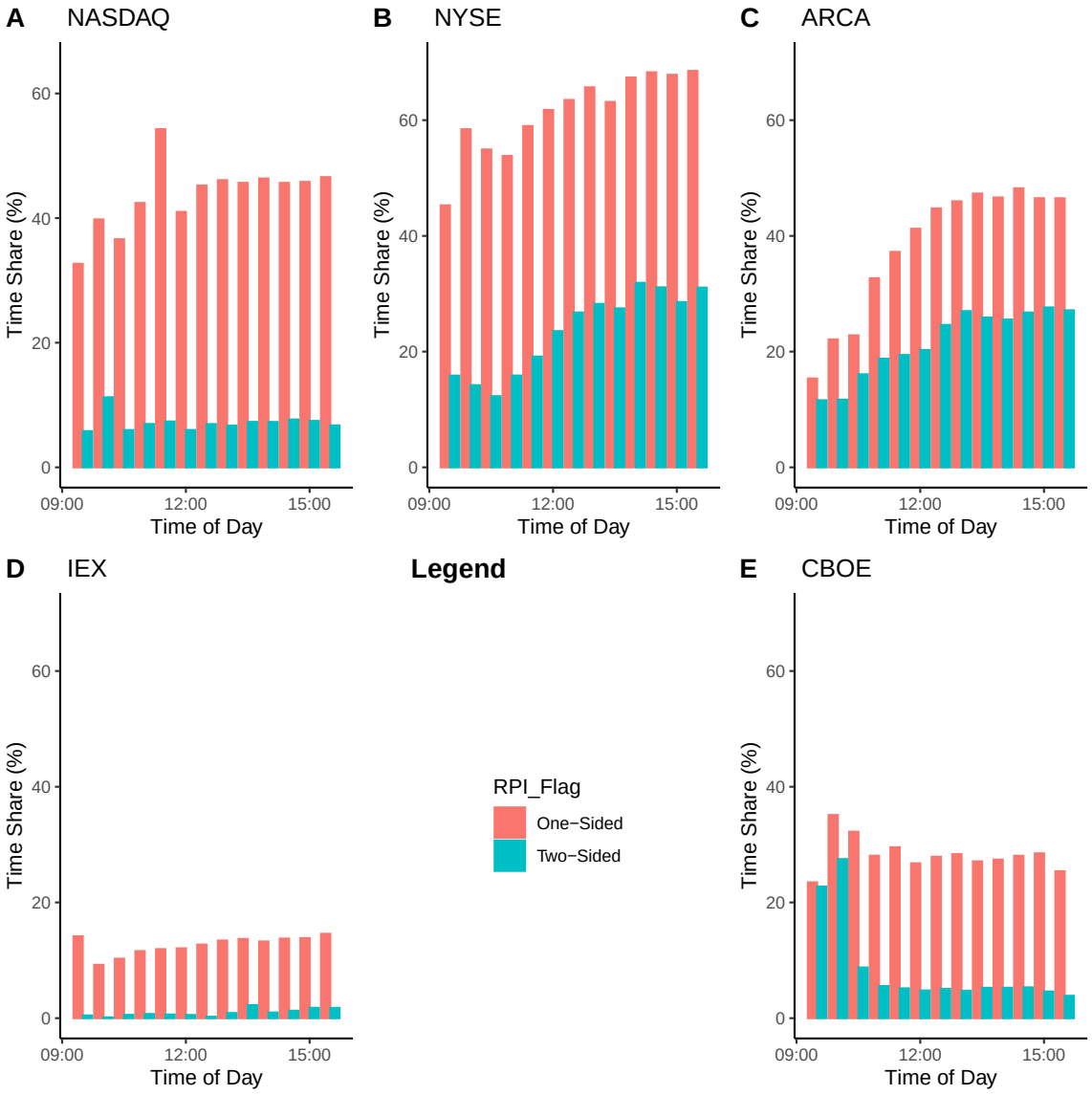
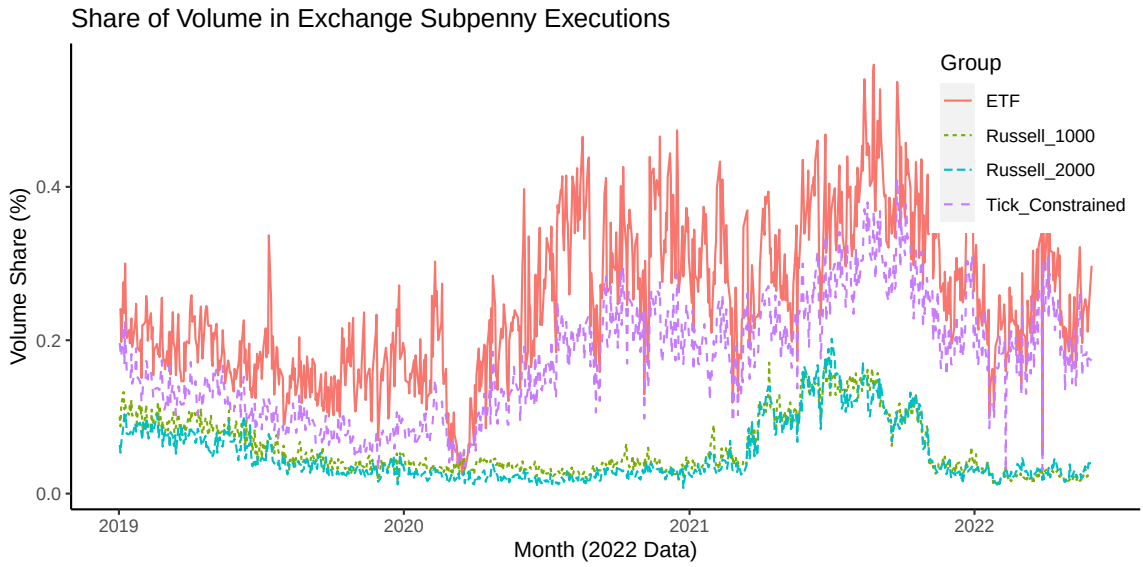
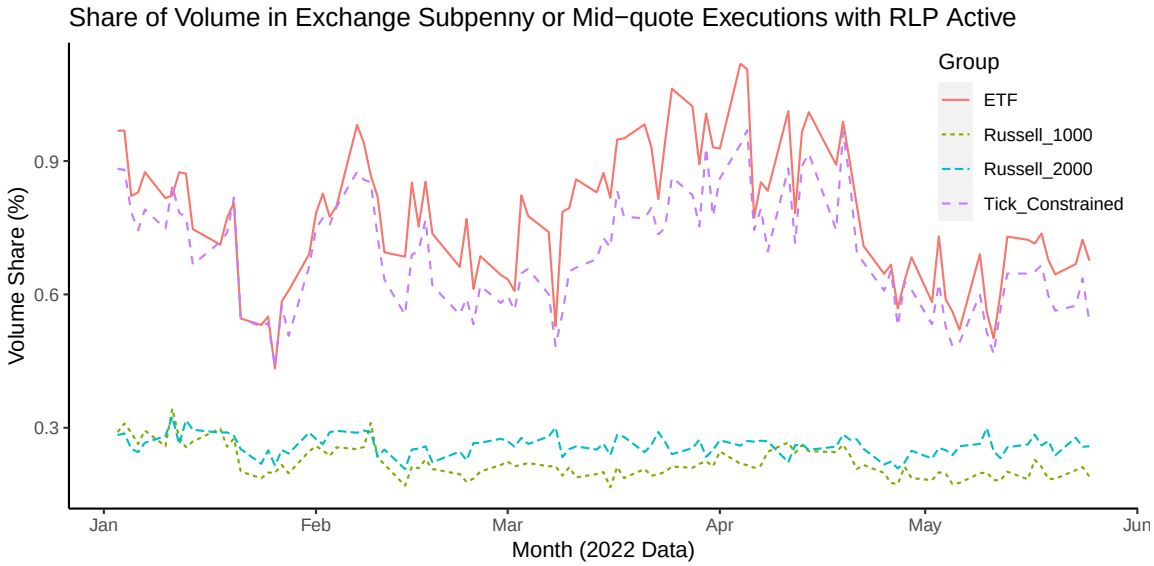


Figure 4. Volume Share of Sub-penny Retail Liquidity Programs. For each day, we plot the share of volume which executes in a retail liquidity program, out of total volume. Our sample can be divided into three groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the day for at least one-third of the days of our sample. Panel A presents the volume share of only exchange sub-penny executions while a RLP indicator is active, while Panel B presents the volume share of all exchange sub-penny or mid-quote executions while a RLP indicator is active.

Panel A: Volume Share for Exchange Sub-penny (Non-Midquote) Executions with an RPI Flag Active



Panel B: Volume Share for Exchange Sub-penny or Mid-quote Executions with an RPI Flag Active



C. *RLP Program Usage and Market Conditions*

Under the current broker's routing structure, market makers must accept order flow from brokers. Market makers do have a choice in where to execute the trade, either by internalizing the trade, or sourcing liquidity from an external venue like an exchange or dark pool. The choice to externalize, however, is not without cost: market makers will fail to capture the spread on any trades they externalize, must pay PFOF if the broker charges PFOF, and will have to pay any trading fees associated with trading on an external venue. For marketable orders, these fees are generally positive. In the model of order-by-order competition, in contrast, bidding is entirely at will. After observing their signal, market makers can post liquidity when they desire to do so, and may withdraw their quotes when they do not.

The current structure of exchange retail liquidity programs has this same at-will feature of liquidity provision, with liquidity-providing participants in the program under no obligation to guarantee execution of retail trades.[13] As a result, exchange retail programs offer insight into the potential workings of an order-by-order model, where market makers are under no obligation to participate for all orders. While many market makers may wish to provide liquidity for orders in large stocks during periods of low volatility, our model suggests this does not hold true for smaller or less liquid stocks. Motivated by this reasoning, we estimate the following regression.

REGRESSION 1: *For each asset i:*

$$\begin{aligned} RPI\_Volume\_Share_i = \quad & \alpha_0 + \alpha_1 Percent\_Time\_At\_Minimum\_Spread_i + \alpha_2 Market\_Cap_i \\ & + \alpha_3 Average\_Volume_i + \epsilon_{ijkt} \end{aligned}$$

Results of Regression 1 are presented in Table II. We estimate volume as a percentage of total volume, and as a percentage of total sub-penny volume. Exchange RLP volume is considerably larger when assets spend a larger percentage of the day at the minimum bid-ask spread, is considerably larger for larger market-cap stocks, and is considerably larger for stocks with higher average trading volume. That small, less liquid stocks have little volume in RLP programs is consistent with the

[13] We note that the NYSE RLP does have a requirement that retail liquidity providers provide price-improving RPI limit orders for at least 5% of the trading day on a certain fraction of trading days to qualify for superior trading fee / rebate pricing. As Figure 1 makes clear, this threshold is low compared to the percentage of time that RPI orders are active.

model prediction that small, less liquid stocks would struggle in the auction format.

Table II: Cross-Sectional Variation in Volume Shares. This table estimates Regression 1 with sub-penny volume, measured as a percentage of all volume, and as a percentage of sub-penny priced volume. For stock i on date t, $Percent\ Time\ At\ Minimum\ Spread_{it}$ measures the percentage of the trading day with a quoted bid-ask spread of one penny, $Volatility_{it}$ measures the standard deviation of 15-minute returns, $Market\ Cap$ measures the market capitalization of the stock in billions, and Average Volume measures the average trading volume in billions. Observations are at the stock (or ETF) level for the sample of securities described in Section VB.

	Dependent variable:	
	Percentage of All Volume (1)	Percentage of Only Sub-penny Volume (2)
Market Cap	0.120***	0.926**
	(0.035)	(0.417)
Percent Time at Minimum Spread	0.001***	0.013***
	(0.0001)	(0.001)
Average Volume	0.041***	0.380***
	(0.004)	(0.042)
Constant	0.058***	0.931***
	(0.002)	(0.030)
Observations	2,590	2,590
R^2	0.159	0.108
Note:		*p<0.1; **p<0.05; ***p<0.01

Retail investors placing market orders may arrive at any time in the day, including during periods of stress. Even under the generous assumption that their orders are uncorrelated with aggregate institutional order flow, half their order flow would be in the same direction as aggregate institutional order flow. To investigate the relationship between retail liquidity program volume and price movements, we estimate Regression 2 with volume and price impacts, with fixed effects for each stock and date, and present the results in Table III. We directly compare on-exchange sub-penny trades with off-exchange sub-penny trades, as these off-exchange trades are the closest analogue to on-exchange trades. Barardehi, Bernhardt, Da, and Warachka (2022) document, however, that sub-penny trading may be driven not by the activity of retail investors, but the extent to which better improvement opportunities (such as mid-quote trading) are not available.

REGRESSION 2: *For each asset i on date t:*

$$\begin{aligned} VolumeShare_{it} = \quad & \alpha_0 + \alpha_1 Percent\_Time\_At\_Minimum\_Spread_{it} + \alpha_2 Volatility_{it} \\ & + \alpha_3 Average\_Volume_i + \alpha_4 Absolute\_Intraday\_Return_{it} + X + \epsilon_{it} \end{aligned}$$

Retail liquidity programs offer less price improvement on average than off-exchange wholesalers. Retail liquidity programs average an improvement of around 10% of the spread. Sub-penny off-exchange trades offer an average improvement of roughly 20% of the spread, while Dyhrberg, Shkilko, and Werner (2022) use SEC Rule 605 reports to estimate that wholesalers offer, on average, price improvement of 40% of the spread. Under the pecking-order theory of Menkveld et al. (2017), investors target low-cost-low-immediacy venues first, and if they fail to find liquidity, they access higher-cost-higher-immediacy venues, particularly at times of market stress or volatility. Consistent with this prediction, we find that on-exchange trading in RLP programs is very sensitive to intra-day volatility, with larger volatility being associated with more exchange sub-penny trading. For off-exchange trading, the opposite is true, with larger volatility associated with less off-exchange sub-penny trading.

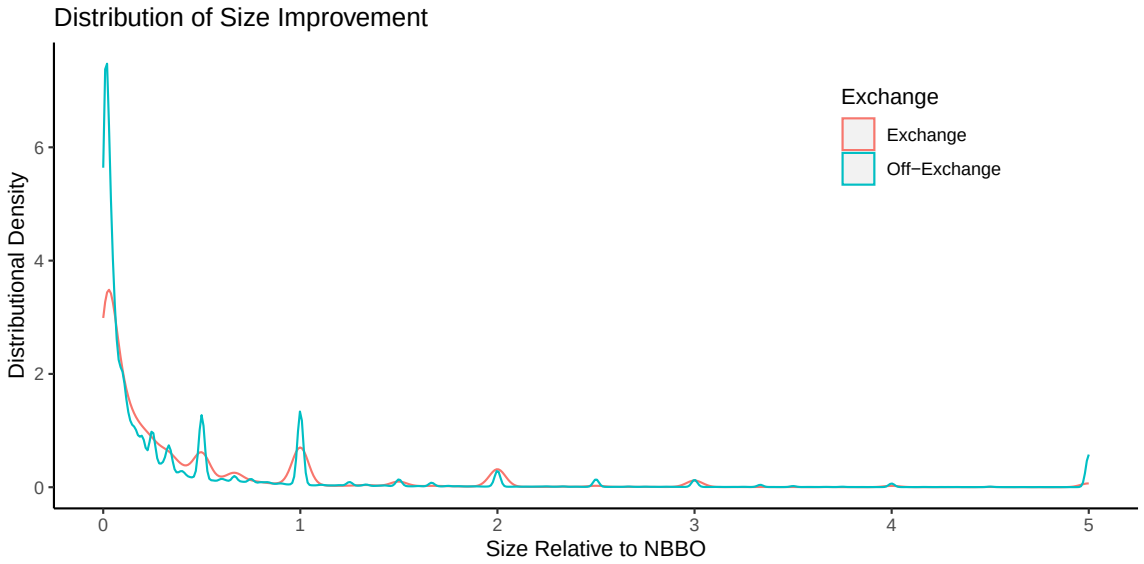
While the Retail Liquidity Programs are the only way that on-exchange trades can be priced in sub-penny increments, retail trades can trade in a variety of methods, with Barber, Huang, Jorion, Odean, and Schwarz (2022) estimating that less than 35% of retail trading takes place at sub-penny prices. Figure 5 depicts the distribution of order sizes for on-exchange and off-exchange sub-penny orders, as a fraction of the NBBO. While a large fraction of sub-penny trades in both venues are odd-lot trades, a far larger share of off-exchange sub-penny trades are for a quantity of shares larger than available at the best bid or offer. Over 2.1% of off-exchange sub-penny trades are for more than 5 times the available shares than the respective national best bid or offer, while only 0.7% of on-exchange sub-penny are for larger than the respective national best bid or offer.

In the economic analysis for the proposed Order-by-Order Competition Rule, the SEC argues that orders with lower price impact are equivalent to lower adverse selection risk: "Marketable orders internalized by wholesalers feature lower price impacts, i.e., have lower adverse selection risk." Securities and Exchange Commission (2022) As one measure of adverse selection, we explore the pattern of order imbalances for on-exchange RLP trades and off-exchange sub-penny trades,

Table III: Panel Variation in Volume and Price Impact. This table estimates Regression 2 with sub-penny volume, expressed as a percentage of total trading volume, and price impact, measured in basis points 30 seconds after trade. Observations are at the stock-day level. Volatility measures the standard deviation of 15-minute price changes. Percent time at Minimum Spread measures the percentage of time the stock spread is a single tick, while absolute intraday return measures the absolute value of the intraday return. We include a fixed effect for each stock and date, and cluster standard errors by stock and by date. Note that Price Impact cannot be calculated when there is zero volume, thus Columns 4, 5, and 6 differ in the number of stock-days with zero volume in each category.

	Dependent Variable:					
	Volume			*Price Impact*		
Venue:	Exchange	Off	Off	Exchange	Off	Off
RPI Active:	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE
	(1)	(2)	(3)	(4)	(5)	(6)
Percent Time At Minimum Spread	0.001	-0.028^{***}	0.027^{***}	0.021	0.050	-0.072^{*}
	(0.010)	(0.005)	(0.009)	(0.021)	(0.038)	(0.037)
Volatility	6.592^{***}	-2.464^{***}	-4.128^{***}	9.520^{**}	2.281^{***}	1.221
	(0.510)	(0.261)	(0.471)	(4.582)	(0.381)	(0.745)
Absolute Intraday Return	1.324^{***}	-0.819^{***}	-0.505^{***}	-0.650	0.251	-0.162
	(0.041)	(0.034)	(0.032)	(0.662)	(0.306)	(0.127)
Observations	1,965,888	1,965,888	1,965,888	682,727	1,771,969	1,885,905
R^2	0.417	0.248	0.380	0.013	0.002	0.003
Residual Std. Error	38.068	27.755	31.363	392.416	403.481	438.930
Note:						*p<0.1; **p<0.05; ***p<0.01

Figure 5. Size Distribution. We plot the distribution of order sizes, as a percentage of the NBBO, for all sub-penny trades occurring in the stocks of our sample on January 3, 2022. We truncate the distribution of orders at 5 times the NBBO. Of all sub-penny trades, 2.1% of all off-exchange sub-penny trades are larger than five times the NBBO, while 0.7% of on-exchange sub-penny trades are larger than five times the NBBO.

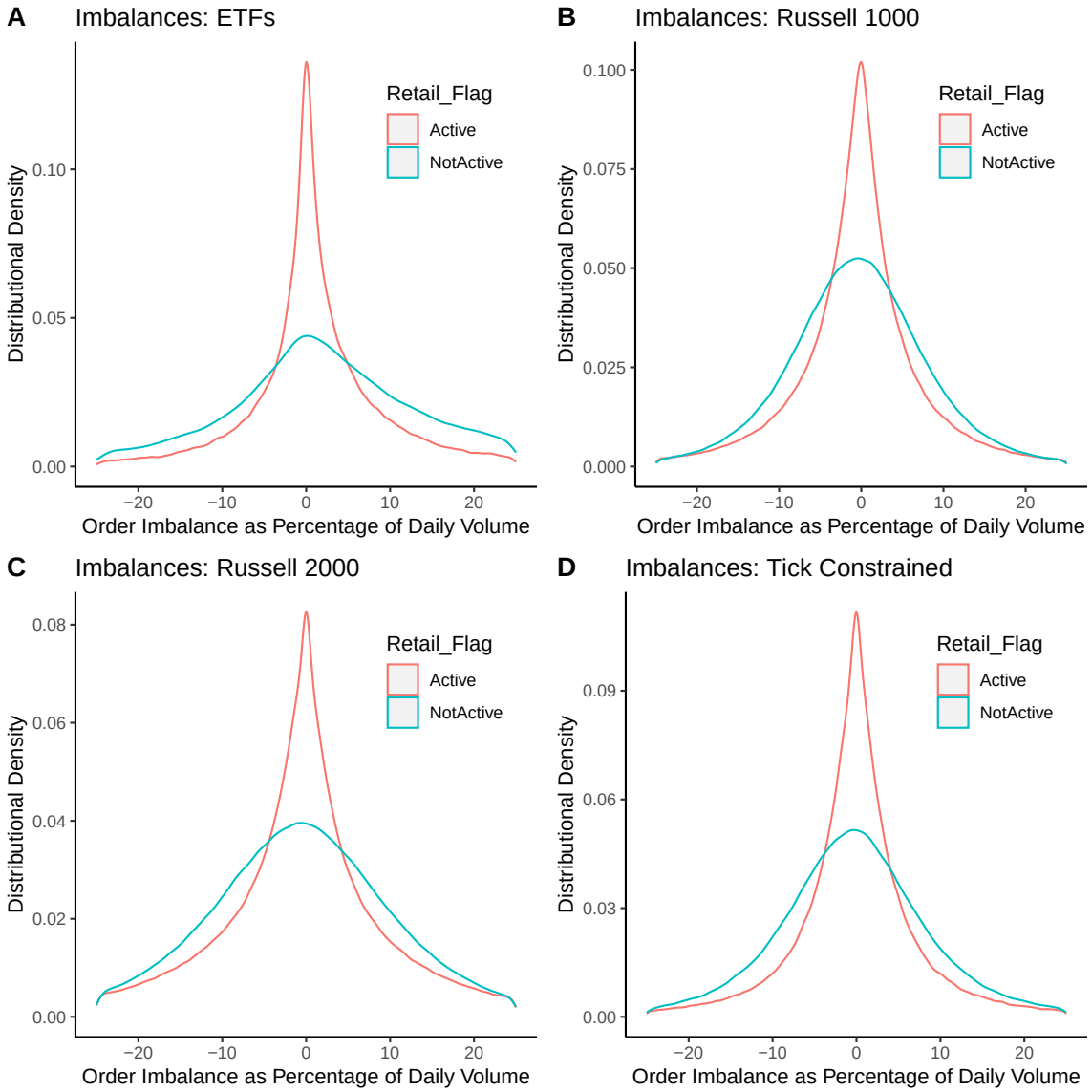


depicted in Figure 6. When the retail flag is active, order imbalances are tightly clustered around a near-zero imbalance, with as many buys arriving as sells. When the retail flag is not active, order imbalances have a distribution with a much larger variance, with a much greater likelihood of large positive or negative order imbalances. This is consistent with the entirely discretionary nature of the RLPs. The SEC views the opportunity for "institutional investors to interact with retail flow" as desirable[14], but it is important to note that institutional investors may be eager to buy from retail investors at times, or sell to retail investors at times, but unlikely to want to stand ready to buy or sell to retail investors at any time on demand.

We also investigate the interaction between RLP trading volume and prior or subsequent quoted bid-ask spreads, both when the RPI Flag is active and inactive. Figure 7 presents the ratio of quoted spreads before and after trades. We first divide trading volume into on-exchange and off-exchange trades, and then further divide volume into sub-penny, mid-quote, and at-quote bins. For each individual stock, we observe the quoted bid-ask spread q_{t+i}, where i can be ± 30 seconds, ± 3 milliseconds, or ± 1 milliseconds. We then calculate the average spread $\bar{q}_{t+i}$ separately for when

[14]See SEC (2013).

Figure 6. Distribution of Order Imbalances. For each stock-day observation, we calculate the total order imbalance among trades occurring when the RPI Flag is active, and the total order imbalance among trades occurring when the RPI Flag is not active. For stock i on date t with flag j, imbalance is calculated as $Imbalance_{ijt} = \frac{\sum Buy_{ijt} - \sum Sell_{ijt}}{\sum Buy_{ijt} + \sum Sell_{ijt}}$. We plot the distribution of imbalances, with the tails truncated to an imbalance of $\pm 50\%$. Panel A presents the imbalance distribution for ETFs, Panel B for stocks in the Russell 1000 Index, Panel C for stocks of the Russell 2000 Index, and Panel D for stocks and ETFs which are tick-constrained, defined as having at a one-penny bid-ask spread least 50% of the trading day for at least one-third of the trading days in our sample.

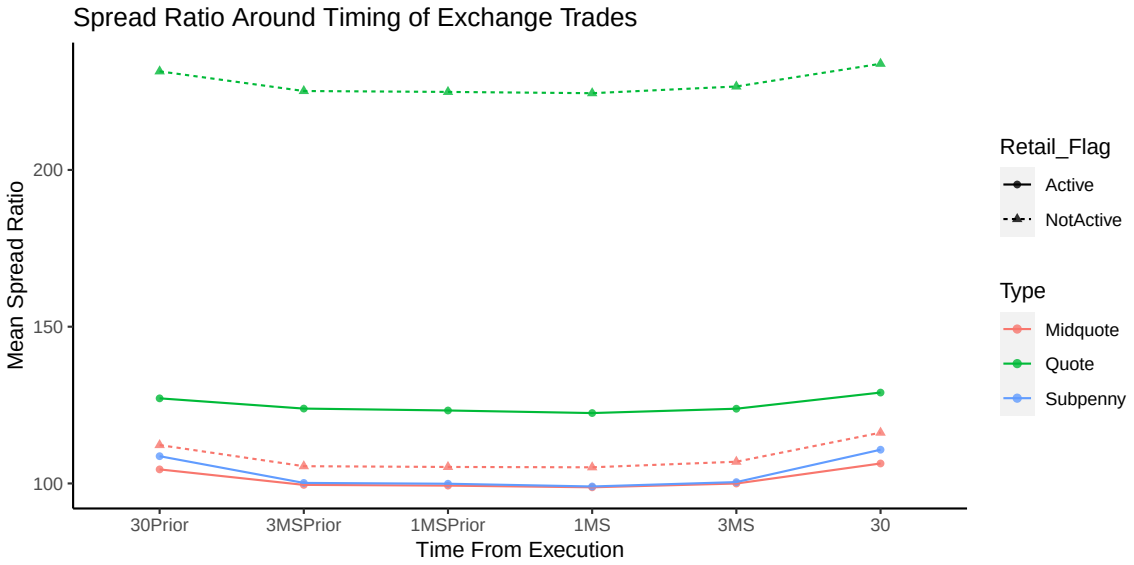


the RPI Flag is or is not active, and plot the ratio $\frac{\bar{q}_i}{\bar{q}}$. When the retail flag is active, off-exchange spreads are very stable, with the same bid-ask spread before and after a trade. When the retail flag is not active, off-exchange spreads before and after a trade tend to be around 2 to 4% wider

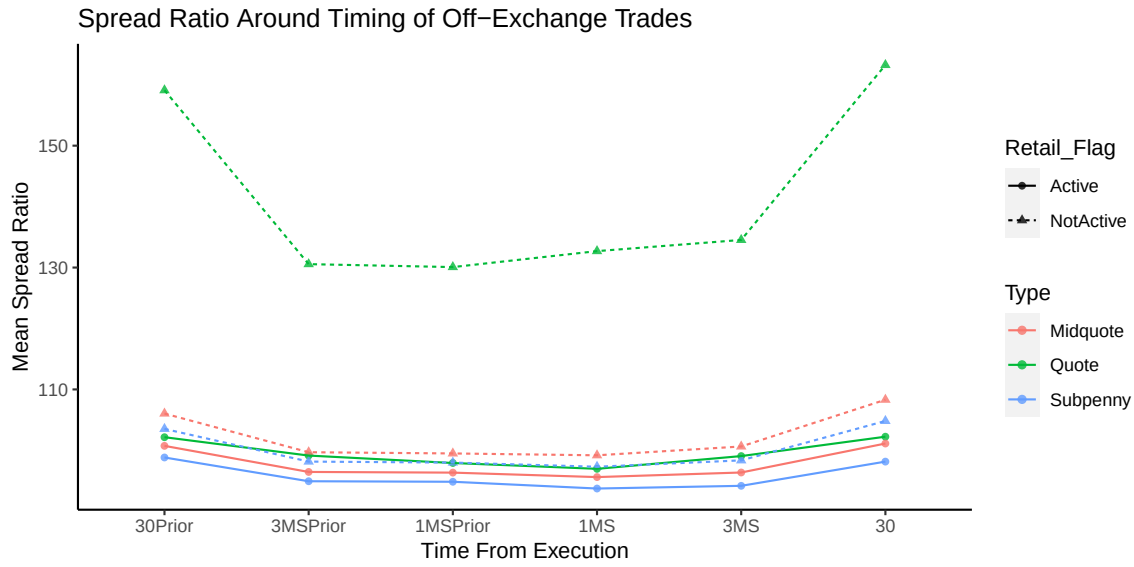
on average, for all categories of pricing. The large discrepancy in quoted spread ratios before and after a trade provides additional suggestive evidence for the pecking order of Menkveld et al. (2017). The discrepancy in spread ratios around the timing of on-exchange mid-quote and sub-penny trades occur at a momentarily liquid time, when quoted spreads are narrow. In contrast with on-exchange trading, the off-exchange trading spread ratios are far more consistent, with the quoted spread width at the time of trades being very similar to the quoted spread width before or after the trade.

Figure 7. Spread Ratios Around Trades. We plot the change in spreads around the timing of a trade, separately for trades occurring when the RPI Flag is active, and trades occurring when it is not active. For trades occurring at time t, we calculate the quoted spread q_t as well as the quoted spread q_{t+i} occurring at a time-offset of i. We then calculate the mean quoted spreads $\bar{q}$ and $\bar{q}_{+i}$, and plot their ratio $r = \frac{\bar{q}_i}{\bar{q}}$. We consider time offsets of 30 seconds prior to trade, 3 milliseconds prior to trade, 1 millisecond prior to trade, 1 millisecond after trade, 3 milliseconds after trade, and 30 seconds after trade.

Panel A: Spread Ratio For Exchange Trades



Panel B: Spread Ratio For Off-Exchange Trades



D. SEC Proposal And Current RLP Usage

The SEC's economic analysis of the Proposed Rule 615 suggests that under the new auction format, institutional traders would give retail traders better trade prices.[15] While the SEC's analysis uses CAT data, the IEX RLP offers an alternative method for estimation of the interest of institutions in trading with retail at mid-quote. The IEX RLP allows market participants to post limit orders priced at the mid-quote which are only available to retail traders.

Figure 3 shows that the IEX RLP has, on average, any interest less than 20% of the trading day. Furthermore, the IEX RLP has two-sided interest less than 5% of the trading day. Figure 8 plots mid-quote trading volume at IEX; total hidden mid-quote orders at IEX (both RLP-only and traditional hidden orders) comprise around 1% to 1.5% of total U.S. equity trading volume, with no obvious change in this volume around the time the IEX RLP is created on October 1, 2019. The IEX RLP began distributing an indicator message when RLP volume is available on October 13, 2021. We note that of the mid-quote volume occurring at IEX, the share of mid-quote orders which are retail orders trading with RLP liquidity is only around 0.05% to 0.10% of total U.S. equities trading volume.[16]

The SEC analysis of CAT data finds that there are many institutional dark orders priced at mid-quote during the time retail investors are active. The suggestion in the SEC economic analysis that these institutional traders will trade with retail at mid-quote in auctions raises the question of why these institutional traders so infrequently seek to trade with retail in the IEX RLP. One possible explanation is that institutions are seeking other large institutions, and do not view the value of trading with retail as worth the risk of information leakage, and switching to auctions would not change the general economics of this calculation.[17] Another possibility is that posting

[15]The SEC reports that "On average, 51% of the shares of individual investor marketable orders internalized by wholesalers are executed at prices less favorable than the NBBO midpoint (Wholesaler Pct Exec Shares Worse Than Midpoint). Out of these individual investors shares that were executed at prices less favorable than the midpoint, on average, 75% of these shares could have hypothetically executed at a better price against the non-displayed liquidity resting at the NBBO midpoint on exchanges and NMS Stock ATSs." Securities and Exchange Commission (2022)

[16]From the TAQ data, it is impossible to determine the exact portion of orders that are retail orders in the IEX RLP program, but we can estimate an upper and lower bound. For the upper bound, we count all mid-quote orders which occur when the IEX retail flag is active, though some of this volume may include non-retail mid-quote orders interacting with hidden mid-quote liquidity. For the lower bound, we measure mid-quote volume which has a simultaneous message update for the RLP program; this measures only retail orders which consume the available RLP liquidity (necessitating an updated RLP message), but will miss retail orders which do not consume all available RLP liquidity and therefore send no update message.

[17]The switch to auctions could potentially make the information leakage problem worse. When trading at mid-quote, no trade direction is identified. In auctions, the trade direction of the incoming retail order would be identified,

in the IEX RLP does not enable trading with retail at mid-quote. We investigate this claim by looking at the distribution of trade prices as a function of the IEX RLP status.

FINRA Rule 5310 requires broker-dealers to route to the best market for a security under prevailing market conditions. To the extent that RLPs offer improvement, wholesalers are already required to route to them; to the extent that RLPs offer inferior price or size improvement, however, wholesalers and brokers would be required to *not* route to them, provided they can obtain favorable price improvement or size improvement off-exchange. In the proposed auctions, wholesalers could internalize orders at mid-quote without routing to an auction. In the current market system, wholesalers can internalize at mid-quote without routing to the IEX RLP.

We investigate whether wholesalers ever fill retail investor orders at prices worse than mid-quote when the IEX RLP has potentially better prices available. We plot the distribution of sub-penny prices for a single trading day in Figure 9. For both on-exchange and off-exchange trades, there is more mid-quote volume when the IEX RPI Flag is active compared to when there is no active RPI Flag, and there is more mid-quote volume when the flag is two-sided (interest in both buying and interest in selling at the mid-quote) than when it is one-sided. While there are off-exchange sub-penny fills at prices worse than mid-quote, Battalio, Jennings, Saglam, and Wu (2022) document that many sub-penny trades are non-retail. For exchange trades, we note that there is precisely zero activity in non-IEX Retail Liquidity Programs when the IEX RLP has two-sided liquidity. Exchange RLP trades are guaranteed to be only retail, so the the complete absence of exchange RLP trades is suggestive evidence that broker-dealers follow FINRA Rule 5310, and route to the IEX RLP if there is active mid-quote interest and are unwilling to either internalize the order at mid-quote or are unable to find an alternative source of mid-quote liquidity.

so that a mid-quote fill would indicate whether the non-retail auction bid was on the buy side or sell side.

Figure 8. IEX Midquote Volume and Key RLP Rule Changes. The IEX Retail Liquidity Program was introduced on October 1, 2019, with only hidden discretionary midpoint-peg orders. On October 13, 2021, the Retail Liquidity Program changed the RLP order type to a midpoint-peg order and began dissemination of an indicator of whether there was RLP interest. On November 22, 2021, the requirement that retail traders submit no more than 390 orders per day was lifted. We plot total midquote volume on IEX (as a percentage of total equities trading volume) with the solid red line. We plot midquote volume which occurs during the time that the IEX RLP is active with the dotted green line. We plot the total midquote volume which occurs simultaneously with an RLP message with the dashed blue line.

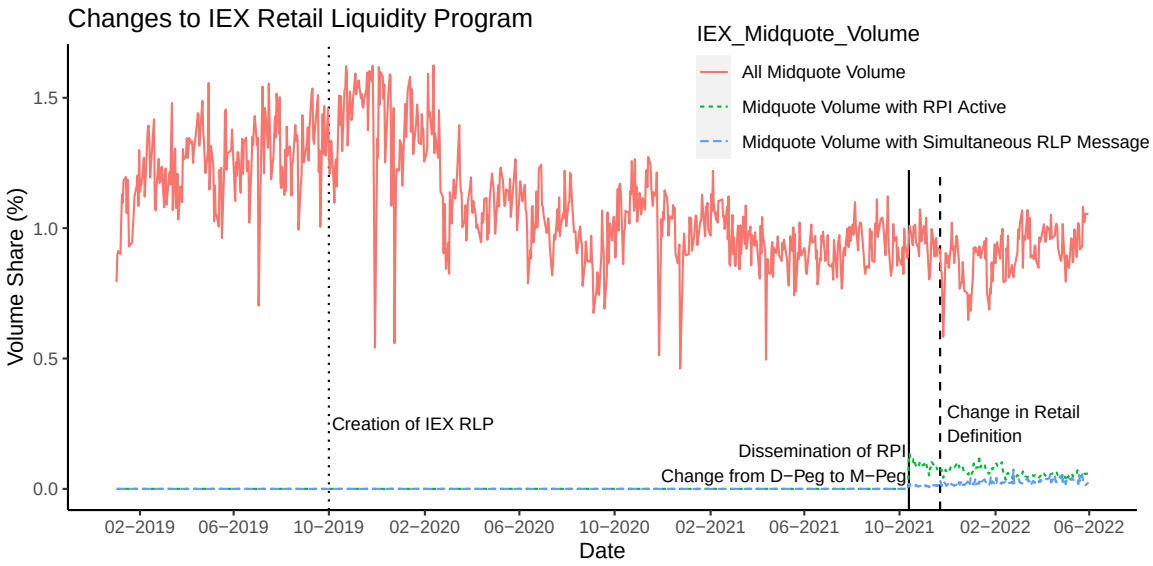
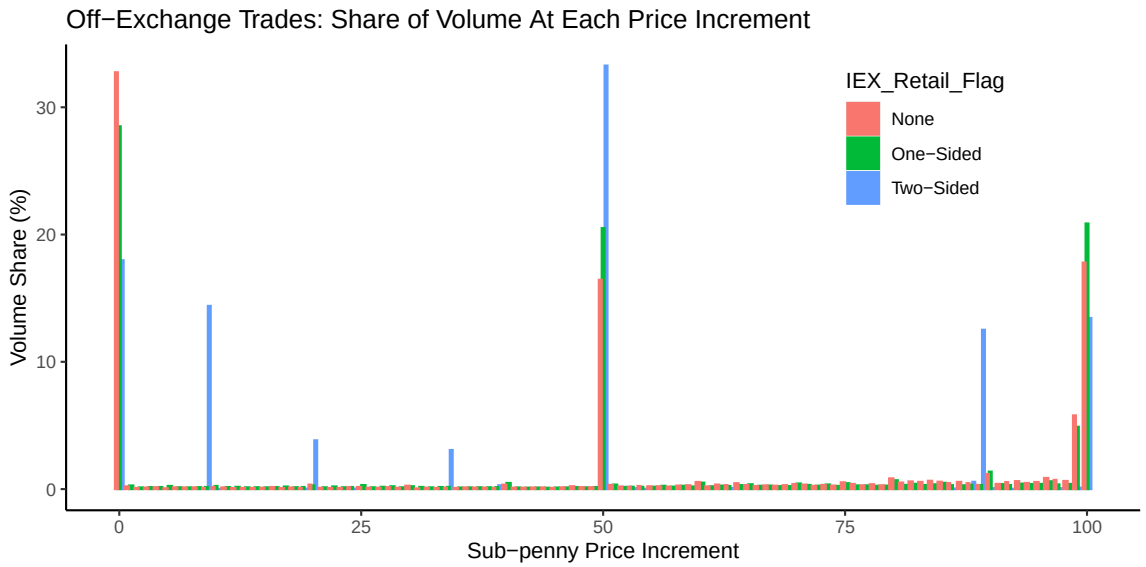
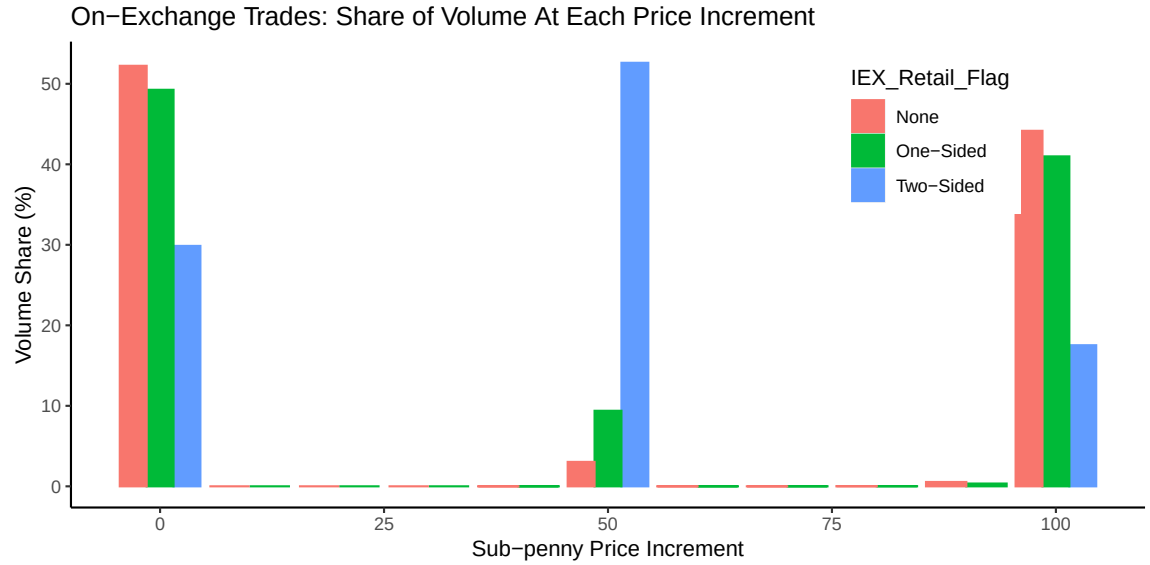


Figure 9. Volume Share of Off-Exchange Sub-Penny Prices. For each possible sub-penny price increment, we plot the volume occurring at this price increment, as a percentage share of all volume occurring on that venue. We use data from trades occurring on January 3, 2022, and separate trade volume into three categories: trades occurring when the IEX RLP has no interest, trades occurring when the IEX RLP has one-sided interest, and trades occurring when the IEX RLP has two-sided interest. Panel A presents the distribution of trades for off-exchange trades. Panel B presents the distribution of trades for on-exchange trades. Note that on-exchange sub-penny trades can only occur in increments of tenths of a cent.

Panel A: Price Improvement Distribution for Off-Exchange Trades



Panel B: Price Improvement Distribution for On-Exchange Trades



V. Conclusion

In the current market structure, retail brokers set up relationships with market makers, and send individual orders to individual market makers. While market makers are evaluated on the aggregate execution quality they deliver, there is no pre-trade communication over individual orders. The SEC concept for order-by-order auctions would require each individual order to be exposed in a bidding process.

Our model shows that a switch to order-by-order auctions comes with trade-offs. Allocative efficiency is improved, as order-by-order auctions ensure that an incoming retail market order is always routed to the market maker who has observed the lowest cost signal. Given the common-value nature of the auction, however, there is a winner's curse. Market makers obtain higher profits in the auction relative to the broker's routing system. Retail investors can be worse off in the switch to order-by-order auctions, particularly in illiquid stocks or at times when interest in voluntary liquidity provision is low, as market participants could opt not to provide any liquidity in the auction.

Our model focuses on inventory cost and competition, and abstracts away from asymmetric information. In bidding in order-by-order auctions, market makers only worry about aggregate inventories. In practice, some market participants bidding in order-by-order auctions may be seeking to trade directionally based on asset price information; this behavior would amplify the winner's curse problem in auctions. We also leave out a consideration of the trade correction and execution guarantees that market makers provide to brokers, which order-by-order auctions would not have.

We empirically evaluate Retail Liquidity Programs (RLPs) to gain insight into how an order-by-order auction would function. Much like the proposed order-by-order auctions, these RLPs allow any market participant to bid potential price improvement to incoming retail market orders. While these RLPs offer potential price improving liquidity, this liquidity is very rarely offered in less liquid stocks, and disappears in times of volatility. As in our theoretical model of order-by-order auctions, observed trades in RLP programs tend to occur at times of lower volatility, on one side of the market, and times when order imbalances are smaller.

REFERENCES

Baldauf, Markus, Joshua Mollner, and Bart Z. Yueshen, 2022, Siphoned Apart: A Portfolio Perspective on Order Flow Fragmentation, *Available at SSRN 4173362* .

Barardehi, Yashar, Dan Bernhardt, Zhi Da, and Mitch Warachka, 2022, Institutional Liquidity Demand and the Internalization of Retail Order Flow: The Tail Does Not Wag the Dog .

Barber, Brad M, Xing Huang, Philippe Jorion, Terrance Odean, and Christopher Schwarz, 2022, A (Sub) penny For Your Thoughts: Tracking Retail Investor Activity in TAQ, *Available at SSRN 4202874* .

Bartlett, Robert P, Justin McCrary, and Maureen O'Hara, 2022, A Fractional Solution to a Stock Market Mystery, *Available at SSRN* .

Battalio, Robert, and Craig W Holden, 2001, A Simple Model of Payment for Order Flow, Internalization, and Total Trading Cost, *Journal of Financial Markets* 4, 33–71.

Battalio, Robert, and Robert Jennings, 2022, Why do Brokers who do not Charge Payment for Order Flow Route Marketable Orders to Wholesalers?, Technical report, Working Paper.

Battalio, Robert, Robert Jennings, Mehmet Saglam, and Jun Wu, 2022, Identifying Market Maker Trades as "Retail" From TAQ: No Shortage of False Negatives and False Positives .

Battalio, Robert, Robert Jennings, and Jamie Selway, 2001, The Relationship Among Market-making Revenue, Payment for Order Flow, and Trading Costs for Market Orders, *Journal of Financial Services Research* 19, 39–56.

Bernhardt, Dan, and Eric Hughson, 1997, Splitting orders, *The Review of Financial Studies* 10, 69–101.

Biais, Bruno, David Martimort, and Jean-Charles Rochet, 2000, Competing mechanisms in a common value environment, *Econometrica* 68, 799–837.

Bishop, Allison, 2022, "The SEC Isn't Mad at PFOF. They're Just Disappointed.", Proof Trading, [Accessed: 2022 02 01].

Bryzgalova, Svetlana, Anna Pavlova, and Taisiya Sikorskaya, 2022, Retail Trading in Options and the Rise of the Big Three Wholesalers, *Available at SSRN* .

Comerton-Forde, Carole, Katya Malinova, and Andreas Park, 2018, Regulating dark trading: Order flow segmentation and market quality, *Journal of Financial Economics* 130, 347–366.

Corwin, Shane A, and Jay F Coughenour, 2008, Limited attention and the allocation of effort in securities trading, *The Journal of Finance* 63, 3031–3067.

Dyhrberg, Anne Haubo, Andriy Shkilko, and Ingrid M Werner, 2022, The Retail Execution Quality Landscape, *Fisher College of Business Working Paper* 014.

Eaton, Gregory W, T Clifton Green, Brian S Roseman, and Yanbin Wu, 2022, Retail Trader Sophistication and Stock Market Quality: Evidence from brokerage outages, *Journal of Financial Economics* 146, 502–528.

Ernst, Thomas, and Chester S Spatt, 2022, Payment for Order Flow and Asset Choice, *NBER Working Paper 29883* .

Foley, Sean, Anqi Liu, Katya Malinova, Andreas Park, and Andriy Shkilko, 2020, Cross-Subsidizing Liquidity, Technical report, Working Paper, Macquarie University.

Gensler, Gary, 2021, "Prepared Remarks at the Global Exchange and FinTech Conference", Speech: Prepared Remarks at the Global Exchange and FinTech Conference [Accessed: 2022 08 29].

Gensler, Gary, 2022, "Competition and the Two SECs:", Remarks Before the SIFMA Annual Meeting, Washington, D.C. [Accessed: 2022 11 01].

Glosten, Lawrence R, and Paul R Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of financial economics* 14, 71–100.

Hendershott, Terrence, Saad Ali Khan, and Ryan Riordan, 2022, Option Auctions, *Working Paper* .

Hu, Edwin, and Dermot Murphy, 2022, Competition for Retail Order Flow and Market Quality, *Available at SSRN 4070056* .

Jain, Pankaj K, Jared A Linna, and Thomas H McInish, 2021, An examination of the NYSE's retail liquidity program, *The Quarterly Review of Economics and Finance* 80, 367–373.

Jain, Pankaj K, Suchi Mishra, Shawn O'Donoghue, and Le Zhao, 2020, Trading Volume Shares and Market Quality in a Zero Commission World, *Available at SSRN 3741470* .

Klemperer, Paul, 1999, Auction theory: A guide to the literature, *Journal of economic surveys* 13, 227–286.

Kyle, Albert S, 1985, Continuous auctions and insider trading, *Econometrica: Journal of the Econometric Society* 1315–1335.

Menezes, Flavio M, and Paulo K Monteiro, 2004, *An introduction to auction theory* (OUP Oxford).

Menkveld, Albert J, Bart Zhou Yueshen, and Haoxiang Zhu, 2017, Shades of darkness: A pecking order of trading venues, *Journal of Financial Economics* 124, 503–534.

Parlour, Christine A., and Uday Rajan, 2003, Payment for Order Flow, *Journal of Financial Economics* 68, 379–411.

Schwarz, Christopher, Brad M Barber, Xing Huang, Philippe Jorion, and Terrance Odean, 2022, The 'Actual Retail Price' of Equity Trades, *Available at SSRN 4189239* .

Schwenk-Nebbe, Sander, 2021, The Participant Timestamp: Get The Most Out Of TAQ Data, *Available at SSRN 3984827* .

SEC, 2013, Self-Regulatory Organizations; The NASDAQ Stock Market LLC; Order Granting Approval to Proposed Rule Change, as Modified by Amendment No. 1, to Establish the Retail Price Improvement Program on a Pilot Basis until 12 Months from the Date of Implementation.

Securities and Exchange Commission, 2022, "Order Competition Rule", Proposal for Rule 615. [Accessed: 2022 12 16].

Appendix A. A microfoundation of inventory cost structure ζ_i

In our baseline model, we assume that the marginal inventory cost for market maker i to execute a sell order is

$$\zeta_i = c_0 + c_1 \frac{1}{N} \sum_{j=1}^{N} y_j + c_2 y_i.$$

In this section, we provide a microfoundation of this formulation, and illustrate the relations between stock liquidity and cost parameters c_0, c_1 and c_2.

We consider a one-period framework, with time $t = 0, 1$. There are N market makers labeled by $i = 1, 2..N$. For any market maker i, if its total net long position is z_i from time 0 to time 1, then it incurs a total inventory cost

$$\frac{1}{2}\gamma z_i^2$$

during this time period, and thus the marginal cost of executing the sell order is γz_i.[18] We assume that at the beginning of time 0, each market maker i's net long position is y_i. The sell order is then assigned to one of the N market makers according to a trading mechanism (broker's routing or order-by-order auctions). If market maker i obtains the sell order, it has to execute the order by internalizing it, routing it to other market makers (inter-dealer market), or sending it to the exchange. Right after market maker i receives the sell order, with probability $\alpha \in (0, 1)$, an inventory shock arrives, the market maker can not internalize the order, and has to either send the order to the exchange or execute it through the inter-dealer market. With probability η, there is active trading of the stock on the exchange, and market maker i can send the order to the exchange and close the position at cost $\bar{s}$. With probability $(1 - \eta)$, the market maker i can only send the order (randomly) to another market maker j. In this case, the cost is

$$\gamma_0 + \gamma y_j$$

where γ_0 is the fixed cost of connecting to another market maker and γy_j is the price charged by market maker j. For simplicity, we assume that market maker j offers competitive price γy_j which is its marginal inventory cost. For simplicity, we make two implicit assumptions here. First, $\bar{s}$ is large enough, so it's always optimal for the market maker to internalize the order when the

[18] This quadratic cost structure is commonly used in the literature (eg. Baldauf Mollner and Yuezhen 2022).

inventory shock is absent, and second, γ_0 is large enough so it's always optimal for the market maker to send the order to the exchange but not other market makers if possible.

Then the expected (marginal) cost of market maker i obtaining the sell order is

$$(1-\alpha)\gamma y_i + \alpha\left[\eta\bar{s} + (1-\eta)\frac{1}{N-1}\sum_{j\neq i}(\gamma_0 + \gamma y_j)\right].$$

The above cost can be rewritten as

$$[\alpha\eta\bar{s} + (1-\eta)\gamma_0] + \frac{1}{N}\left(\frac{\alpha(1-\eta)\gamma N}{N-1}\right)\sum_j y_j + \left((1-\alpha)\gamma - \frac{\alpha(1-\eta)\gamma}{N-1}\right)y_i.$$

Let

$$c_0 = \alpha\eta\bar{s} + (1-\eta)\gamma_0,$$

$$c_1 = \frac{\alpha(1-\eta)\gamma N}{N-1},$$

and

$$c_2 = (1-\alpha)\gamma - \frac{\alpha(1-\eta)\gamma}{N-1},$$

then the marginal cost for market maker i to execute the sell order is

$$c_0 + c_1\frac{1}{N}\sum_{j=1}^{N} y_j + c_2 y_i.$$

Stock liquidity is linked to the parameter η in our microfoundation. When a stock is more liquid, it's more likely to have active trading on the exchange at that moment, and thus η will be larger. As a result, the ratio

$$\frac{c_2}{c_1} = \frac{N-1}{\alpha\gamma N}\left((1-\alpha)\frac{\gamma}{1-\eta} - \frac{\alpha\gamma}{N-1}\right)$$

will be larger. We utilize this interpretation in our discussions of model implications.

Appendix B. Proofs

Proof of Proposition 1

Consider any $i \in \{1, 2...N\}$ and $(x, y) \in \left[-\frac{1}{2}, \frac{1}{2}\right] \times \left[-\frac{1}{2}, \frac{1}{2}\right]$, let

$$G(x|y) = \text{Prob}\left[\min_{-i} y_{-i} \leq x | y_i = y\right]$$

and

$$g(x|y) = \frac{dG(x|y)}{dx}.$$

It's easy to show

$$G(x|y) = 1 - \left(\frac{1}{2} - x\right)^{N-1},$$

$$g(x|y) = (N-1)\left(\frac{1}{2} - x\right)^{N-2}.$$

Let

$$\begin{aligned} v(x, y) &= \mathbb{E}\left[c_i | \min_{-i} y_{-i} = x, y_i = y\right] \\ &= c_0 + c_1 \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} y_j | \min_{-i} y_{-i} = x, y_i = y\right] + c_2 \mathbb{E}\left[y_i | \min_{-i} y_{-i} = x, y_i = y\right] \\ &= c_0 + \left(\frac{c_1}{N} + c_2\right) y + \frac{c_1}{N} x + c_1 \frac{N-2}{N} \frac{1}{2}\left(\frac{1}{2} + x\right) \\ &= \left(c_0 + c_1 \frac{N-2}{4N}\right) + \frac{c_1}{2} x + \left(\frac{c_1}{N} + c_2\right) y. \end{aligned}$$

We focus on symmetric equilibria. Suppose market maker i's opponents use a continuous, increasing strategy $\beta(y)$ at time 0. And suppose market maker i observes signal y and reports signal z, its expected profit is

$$\begin{aligned} U_i(z, y) &= \text{Prob}\left(z \leq \min_{-i} y_{-i} | y\right)\left[\beta(z) - \mathbb{E}\left(c | z \leq \min_{-i} y_{-i}, y_i = y\right)\right] \\ &= \left[1 - G(z|y)\right]\left[\beta(z) - \frac{1}{1 - G(z|y)} \int_z^{\frac{1}{2}} g(x|y)\, v(x, y)\, dx\right] \\ &= \left[1 - G(z|y)\right]\beta(z) - \int_z^{\frac{1}{2}} g(x|y)\, v(x, y)\, dx. \end{aligned}$$

Market maker i's optimization condition (necessary condition) is

$$\left.\frac{\partial U_i(z,y)}{\partial z}\right|_{z=y}=0.$$

This is

$$-g(y|y)\beta(y)+(1-G(y|y))\beta'(y)+g(y|y)v(y|y)=0.$$

Simplifying the condition, we get

$$-\beta(y)+\left(\frac{1-G(y|y)}{g(y|y)}\right)\beta'(y)+v(y|y)=0. \tag{B1}$$

Let's conjecture that $\beta(y)$ is linear, i.e., there exist k_0, k_1 such that

$$\beta(y)=k_0+k_1y.$$

Substitute this into (B1), we have

$$-(k_0+k_1y)+\frac{\frac{1}{2}-y}{N-1}k_1+\left(c_0+c_1\frac{N-2}{4N}\right)+\frac{c_1}{2}y+\left(\frac{c_1}{N}+c_2\right)y=0.$$

Then k_0,k_1 are solved by

$$-k_0+\frac{\frac{1}{2}k_1}{N-1}+c_0+c_1\frac{N-2}{4N}=0,$$

$$-k_1-\frac{k_1}{N-1}+\frac{c_1}{2}+\frac{c_1}{N}+c_2=0.$$

Then we get

$$k_1=\frac{N-1}{N}\left(\frac{c_1}{2}\frac{N+2}{N}+c_2\right),$$

$$k_0=c_0+\frac{c_1}{4N}\left(N-1+\frac{2}{N}\right)+\frac{c_2}{2N}.$$

It's easy to check that

$$\left.\frac{\partial U_i(z,y)}{\partial z}\right|_{z=y}=0$$

is also the sufficient condition in the optimization problem in this linear equilibrium because of the linearity of the equilibrium.

Proof of Proposition 2

Consider any $i \in \{1, 2...N\}$ and $(x, w) \in \left[-\frac{1}{2}, \frac{1}{2}\right] \times \left[-\frac{1}{2}, \frac{1}{2}\right]$, let

$$G_0(x|w) = \text{Prob}\left[\min_{-i} w_{-i} \leq x | w_i = w\right]$$

and

$$g_0(x|w) = \frac{dG(x|w)}{dx}.$$

It's easy to show

$$G(x|w) = 1 - \left(\frac{1}{2} - x\right)^{N-1},$$

$$g(x|w) = (N-1)\left(\frac{1}{2} - x\right)^{N-2}.$$

Let

$$\begin{aligned}
v(x, w) =& \mathbb{E}\left[\zeta_i | \min_{-i} w_{-i} = x; w_i = w\right] \\
=& p_0 \mathbb{E}\left[\zeta_i | \min_{-i} w_{-i} = x; w_i = w; \exists j \neq i, w_j = y_j = x\right] \\
&+ (1 - p_0) p_0 \mathbb{E}\left[c_i | \min_{-i} w_{-i} = x; w_i = y; \nexists j \neq i, w_j = y_j = x\right] \\
=& p_0 \left[c_0 + c_1 \left(\frac{x + (N-2) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
&+ (1 - p_0) \left[c_0 + c_1 \left(\frac{(N-1) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
=& c_0 + c_1 \left(\frac{\left(p_0 x + (1 - p_0) p_0 \frac{\frac{1}{2}+x}{2}\right) + (N-2) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w.
\end{aligned}$$

We focus on symmetric equilibria. Suppose all of market maker i's opponents use a continuous, increasing bid strategy

$$B(w) = K_0 + K_1 w$$

at time 0. When market maker i observes signal w and reports signal z, its expected profit is

$$\begin{aligned} U_i(z,w) &= \text{Prob}\left(z \leq \min_{-i} w_{-i} | w_i = w\right)\left[B(z) - \mathbb{E}\left(\zeta_i | z \leq \min_{-i} w_{-i}, w_i = w\right)\right] \\ &= \left[1 - G(z|w)\right]\left[B(z) - \frac{1}{1-G(z|w)}\int_z^1 g(x|w)\, v(x,w)\, dx\right] \\ &= \left[1 - G(z|w)\right] B(z) - \int_z^1 g(x|w)\, v(x,w)\, dx. \end{aligned}$$

Market maker i's marginal incentive is characterized by

$$\begin{aligned} \frac{\partial U_i(z,w)}{\partial z} &= -g(z|w)\, B(z) + (1 - G(z|w))\, B'(z) + g(z|w)\, v(z|w) \\ &= g(z|w)\left[-B(z) + \left(\frac{1-G(z|w)}{g(z|w)}\right) B'(z) + v(z|w)\right] \\ &= g(z|w)\left[\begin{array}{c} -K_0 - K_1 z + \frac{\frac{1}{2}-z}{N-1} K_1 + c_0 \\ +c_1\left(\frac{\left(p_0 z + (1-p_0) p_0 \frac{\frac{1}{2}+z}{2}\right) + (N-2) p_0 \frac{\frac{1}{2}+z}{2} + p_0 w}{N}\right) + c_2 p_0 w \end{array}\right]. \end{aligned}$$

Let's conjecture that in equilibrium we have

$$\left.\frac{\partial U_i(z,w)}{\partial z}\right|_{z=w} = 0. \tag{B2}$$

This implies

$$-(K_0 + K_1 w) + \frac{\frac{1}{2} - w}{N-1} K_1 + c_0 + c_1\left(\frac{\left(p_0 w + (1-p_0) p_0 \frac{\frac{1}{2}+w}{2}\right) + (N-2) p_0 \frac{\frac{1}{2}+w}{2} + p_0 w}{N}\right) + c_2 p_0 w = 0.$$

Since the above condition holds for all w, then K_0,K_1 are solved by

$$-K_0 + \frac{\frac{1}{2} K_1}{N-1} + c_0 + c_1 \frac{N-2}{4N} p_0 + \frac{c_1 (1-p_0) p_0}{4N} = 0,$$

$$-K_1 - \frac{K_1}{N-1} + c_2 p_0 + \frac{2 c_1 p_0}{N} + \frac{c_1 (N-2) p_0}{2N} + \frac{c_1 (1-p_0) p_0}{2N} = 0.$$

Then we get

$$K_1 = \frac{N-1}{N}\left(c_2 p_0 + \frac{2c_1 p_0}{N} + \frac{c_1\left(N-2\right)p_0}{2N} + \frac{c_1\left(1-p_0\right)p_0}{2N}\right), \tag{B3}$$

$$K_0 = c_0 + \frac{p_0}{4N^2}\left[\left(3+N^2-p_0-Np_0\right)c_1 + 2Nc_2\right]. \tag{B4}$$

We also need to verify that condition (B2) is a sufficient condition for optimization. Note that $g\left(z|w\right) > 0$ and

$$-K_0 - K_1 z + \frac{\frac{1}{2}-z}{N-1}K_1 + c_0 + c_1\left(\frac{\left(p_0 z + \left(1-p_0\right)p_0\frac{\frac{1}{2}+z}{2}\right) + \left(N-2\right)p_0\frac{\frac{1}{2}+z}{2} + p_0 w}{N}\right) + c_2 p_0 w$$

is linear in z, then it's clear that with (B3) and (B4), we must have that for all w,

$$\frac{\partial U_i\left(z,w\right)}{\partial z} < 0 \Longleftrightarrow z > w,$$

confirming that (B2) is a sufficient condition for optimization.

Proof of Lemma 1

First let's introduce the random variable

$$r = \min_i y_i \in \left[-\frac{1}{2}, \frac{1}{2}\right]$$

and its CDF

$$H\left(r\right) = 1 - \left(\frac{1}{2} - r\right)^N$$

and PDF

$$h\left(r\right) = N\left(\frac{1}{2} - r\right)^{N-1}.$$

Then the total expected profit of market makers is

$$\begin{aligned}
W_M^{OBO} =& \mathbb{E}\left\{\mathbb{E}\left[k_0 + k_1 r - c_0 - c_1 \frac{1}{N}\sum_{j=1}^{N} y_j - c_2 r | \min_i y_i = r\right]\right\} \\
=& \mathbb{E}\left\{k_0 + k_1 r - c_0 - c_1\left(\frac{r + (N-1)\left(\frac{1}{2}+r\right)\frac{1}{2}}{N}\right) - c_2 r\right\} \\
=& \mathbb{E}\left\{\frac{\left(\frac{1}{2}-r\right)(c_1 + c_2 N)}{N^2}\right\} \\
=& \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\left(\frac{1}{2}-r\right)(c_1 + c_2 N)}{N^2} N\left(\frac{1}{2}-r\right)^{N-1} dr \\
=& \frac{1}{N+1}\left(\frac{c_1}{N} + c_2\right).
\end{aligned}$$

The expected total profit of investors W_I is

$$\begin{aligned}
W_I^{OBO} &= -\mathbb{E}\left[k_0 + k_1 r | \min_i y_i = r\right] \\
&= -\int_{-\frac{1}{2}}^{\frac{1}{2}} (k_0 + k_1 r) N\left(\frac{1}{2}-r\right)^{N-1} dr \\
&= -\left[c_0 + \frac{1}{N(N+1)} c_1 - \frac{N-3}{2(N+1)} c_2\right]
\end{aligned}$$

and the total welfare W_{total} is

$$\begin{aligned}
W_{total}^{OBO} &= \mathbb{E}\left\{\mathbb{E}\left[-c_0 - c_1 \frac{1}{N}\sum_{j=1}^{N} y_j - c_2 r | \min_i y_i = r\right]\right\} \\
&= W_M^{OBO} + W_I^{OBO} \\
&= -\left(c_0 - \frac{N-1}{N+1}\frac{c_2}{2}\right).
\end{aligned}$$

Proof of Lemma 2

let's introduce the random variable

$$r = \min_i w_i \in \left[-\frac{1}{2}, \frac{1}{2}\right]$$

and its CDF

$$H(r) = 1 - \left(\frac{1}{2} - r\right)^N$$

and PDF

$$h(r) = N\left(\frac{1}{2} - r\right)^{N-1}.$$

Then the total expected profit of market makers is

$$\begin{aligned}
W_M^{BR} =& \mathbb{E}\left\{\mathbb{E}\left[t(r) - c_0 - c_1\frac{1}{N}\sum_{j=1}^{N} y_j - c_2 y_i | w_i = \min_j w_j = r\right]\right\} \\
=& \mathbb{E}\left\{K_0 + K_1 r - c_0 - c_1\left(\frac{p_0 r + (N-1)\, p_0\left(\frac{1}{2} + r\right)\frac{1}{2}}{N}\right) - c_2 p_0 r\right\} \\
=& \mathbb{E}\left\{\frac{p_0}{4N^2}\left[2Nc_2(1-2r) + c_1(3-p_0)(1-2r) + c_1 N(1-p_0)(1+2r)\right]\right\} \\
=& \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{p_0}{4N^2}\left[2Nc_2(1-2r) + c_1(3-p_0)(1-2r) + c_1 N(1-p_0)(1+2r)\right] N\left(\frac{1}{2} - r\right)^{N-1} dr \\
=& \frac{p_0(2c_1 - p_0 c_1 + Nc_2)}{N(1+N)}.
\end{aligned}$$

The expected total profit of investors W_I is

$$\begin{aligned}
W_I^{BR} &= -\mathbb{E}\left[K_0 + K_1 r | \min_i w_i = r\right] \\
&= -\int_{-\frac{1}{2}}^{\frac{1}{2}} (K_0 + K_1 r) N\left(\frac{1}{2} - r\right)^{N-1} dr \\
&= -\left[c_0 + p_0\frac{2(2-p_0)c_1 - (N-3)Nc_2}{2N(1+N)}\right],
\end{aligned}$$

and the total welfare W_{total} is

$$\begin{aligned}
W_{total}^{BR} &= \mathbb{E}\left\{\mathbb{E}\left[-c_0 - c_1\frac{1}{N}\sum_{j=1}^{N} y_j - c_2 y_i | w_i = \min_j w_j = r\right]\right\} \\
&= -\left(c_0 - p_0\frac{N-1}{N+1}\frac{c_2}{2}\right).
\end{aligned}$$

Proof of Proposition 3

Both $W_{total}^{BR} < W_{total}^{OBO}$ and $W_{M}^{BR} < W_{M}^{OBO}$ are obvious. And

$$\begin{aligned}
&W_I^{BR} < W_I^{OBO} \\
\Longleftrightarrow &-\left[c_0 + p_0 \frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)}\right] < -\left[c_0 + \frac{1}{N(N+1)}c_1 - \frac{N-3}{2(N+1)}c_2\right] \\
\Longleftrightarrow &\frac{c_2}{c_1} > \frac{2(1-p_0)}{N(N-3)}.
\end{aligned}$$

Proof of Proposition 4

Since wholesalers are not able to observe the realization of $\tilde{c}_0$, they can condition their strategies only on the distributional information about $\tilde{c}_0$. We still focus on symmetric equilibria in this case, and let's conjecture that all wholesalers uses the same bidding strategy

$$\tilde{\beta}\left(y\right) = \tilde{k}_0 + \tilde{k}_1 y,$$

with $\tilde{k}_1 > 0$. Similar to our baseline model, the wholesaler with lowest signal realization obtains the order in equilibrium. We follow the proof of Proposition 2, notably, the function $\tilde{v}\left(x,w\right) = \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} w_{-i} = x; w_i = w\right]$ now becomes

$$\begin{aligned}
\tilde{v}\left(x,w\right) =&\mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} w_{-i} = x; w_i = w\right] \\
=&p_0 \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} w_{-i} = x; w_i = w; \exists j \neq i, w_j = y_j = x\right] \\
&+ \left(1-p_0\right) p_0 \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} w_{-i} = x; w_i = y; \nexists j \neq i, w_j = y_j = x\right] \\
=&p_0 \left[c_0 + c_1 \left(\frac{x + \left(N-2\right) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
&+ \left(1-p_0\right) \left[c_0 + c_1 \left(\frac{\left(N-1\right) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
=&c_0 + c_1 \left(\frac{\left(p_0 x + \left(1-p_0\right) p_0 \frac{\frac{1}{2}+x}{2}\right) + \left(N-2\right) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w,
\end{aligned}$$

which implies that

$$\tilde{v}(x,w) = v(x,w).$$

The rest of the proof follows the proof of Proposition 2. So the equilibrium strategy is the same as that in the baseline model.

Similarly, note that

$$\mathbb{E}(\tilde{c}_0) = c_0,$$

the proof of welfare computation follows our proof of Lemma 2, and thus all welfare outcomes are the same as that in our baseline model.

Proof of Proposition 5

When $\delta_c = 0$, the institutional traders and wholesalers receive i.i.d signals, and they are symmetric. Let's conjecture that all market makers choose the same linear equilibrium strategy

$$\tilde{\beta}_i(y_i; \delta_c = 0) = \tilde{k}_0(\delta_c = 0) + \tilde{k}_1(\delta_c = 0) y_i.$$

The number of market makers is $N + N_0$. We follow the proof of Proposition 1, the function $v(x,y)$ now becomes

$$\begin{aligned}
\tilde{v}(x,y) &= \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} y_{-i} = x, y_i = y\right] \\
&= c_0 + c_1 \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N+N_0} y_j | \min_{-i} y_{-i} = x, y_i = y\right] + c_2 \mathbb{E}\left[y_i | \min_{-i} y_{-i} = x, y_i = y\right] \\
&= c_0 + \left(\frac{c_1}{N+N_0} + c_2\right) y + \frac{c_1}{N+N_0} x + c_1 \frac{N+N_0-2}{N+N_0} \frac{1}{2}\left(\frac{1}{2} + x\right) \\
&= \left(c_0 + c_1 \frac{N+N_0-2}{4(N+N_0)}\right) + \frac{c_1}{2} x + \left(\frac{c_1}{N+N_0} + c_2\right) y,
\end{aligned}$$

which is the $v(x,y)$ function with $(N+N_0)$ wholesalers. For the rest of the proof, we follow the proof of Proposition 1, and we can show that the equilibrium strategy is equivalent to that in Proposition 1 with the number of wholesalers being $N + N_0$.

Proof of Proposition 6

Since the equilibrium of broker's routing is the same as that in the baseline model, we have

$$\tilde{W}_{total}^{BR} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right),$$

$$\tilde{W}_{I}^{BR} = -\left[c_0 + p_0 \frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)}\right],$$

and

$$\tilde{W}_{W}^{BR} = \frac{p_0\left(2c_1 - p_0c_1 + Nc_2\right)}{N\left(1+N\right)}.$$

For order-by-order auctions, the welfare outcomes are

$$\tilde{W}_{total}^{OBO} = -\left(c_0 - \frac{N+N_0-1}{N+N_0+1}\frac{c_2}{2}\right),$$

$$\tilde{W}_{I}^{OBO} = -\left[c_0 + \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)}c_1 - \frac{N+N_0-3}{2\left(N+N_0+1\right)}c_2\right],$$

and

$$\tilde{W}_{W}^{OBO} = \frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{c_1}{N+N_0} + c_2\right).$$

First, it's obvious that

$$\tilde{W}_{total}^{OBO} > \tilde{W}_{total}^{BR},$$

because $p_0 \in (0,1)$ and $N_0 > 1$. Second,

$$\begin{aligned}
&\tilde{W}_{W}^{BR} < \tilde{W}_{W}^{OBO} \\
\Longleftrightarrow &\frac{p_0\left(2c_1 - p_0c_1 + Nc_2\right)}{N\left(1+N\right)} < \frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{c_1}{N+N_0} + c_2\right) \\
\Longleftrightarrow &\frac{p_0\left(2 - p_0 + N\frac{c_2}{c_1}\right)}{N\left(1+N\right)} < \frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{1}{N+N_0} + \frac{c_2}{c_1}\right) \\
\Longleftrightarrow &\frac{p_0\left(2-p_0\right)}{N\left(1+N\right)} + \frac{p_0}{1+N}\frac{c_2}{c_1} < \frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{1}{N+N_0} + \frac{c_2}{c_1}\right) \\
\Longleftrightarrow &\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}\right)\frac{c_2}{c_1} > -\left(\frac{N}{\left(N+N_0\right)^2}\frac{1}{N+N_0+1} - \frac{p_0\left(2-p_0\right)}{N\left(1+N\right)}\right) \\
\Longleftrightarrow &\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}\right)\frac{c_2}{c_1} > -\frac{1}{N+N_0}\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{\left(1+N\right)}\frac{\left(2-p_0\right)\left(N+N_0\right)}{N}\right)
\end{aligned}$$

Since

$$\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}>\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N},$$

we know that

$$\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}\right)\frac{c_2}{c_1}>-\frac{1}{N+N_0}\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)$$

is equivalent to

$$\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}>0\Longleftrightarrow p_0<\frac{N}{N+N_0}\frac{1+N}{N+N_0+1}$$

and

$$\frac{c_2}{c_1}>\frac{-\frac{1}{N+N_0}\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)}{\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}}.$$

Finally,

$$\begin{aligned}
&\tilde{W}_I^{BR}<\tilde{W}_I^{OBO}\\
\Longleftrightarrow&-\left[c_0+p_0\frac{2(2-p_0)c_1-(N-3)Nc_2}{2N(1+N)}\right]<-\left[c_0+\frac{1}{(N+N_0)(N+N_0+1)}c_1-\frac{N+N_0-3}{2(N+N_0+1)}c_2\right]\\
\Longleftrightarrow&p_0\frac{2(2-p_0)c_1-(N-3)Nc_2}{2N(1+N)}>\frac{1}{(N+N_0)(N+N_0+1)}c_1-\frac{N+N_0-3}{2(N+N_0+1)}c_2\\
\Longleftrightarrow&p_0\frac{2(2-p_0)-(N-3)N\frac{c_2}{c_1}}{2N(1+N)}>\frac{1}{(N+N_0)(N+N_0+1)}-\frac{N+N_0-3}{2(N+N_0+1)}\frac{c_2}{c_1}\\
\Longleftrightarrow&\left(\frac{N+N_0-3}{2(N+N_0+1)}-p_0\frac{(N-3)}{2(1+N)}\right)\frac{c_2}{c_1}>\frac{1}{(N+N_0)(N+N_0+1)}-p_0\frac{(2-p_0)}{N(1+N)}\\
\Longleftrightarrow&\frac{c_2}{c_1}>\frac{\frac{1}{(N+N_0)(N+N_0+1)}-\frac{p_0(2-p_0)}{N(1+N)}}{\frac{N+N_0-3}{2(N+N_0+1)}-p_0\frac{(N-3)}{2(1+N)}}.
\end{aligned}$$

The last inequality holds because we always have $\frac{N+N_0-3}{2(N+N_0+1)}-p_0\frac{(N-3)}{2(1+N)}>0$.

Proof of Proposition 7

We need to verify that this is indeed an equilibrium. Let $\underline{\delta}=\max\{\delta_{c1},\delta_{c2}\}$ where δ_{c1} is defined by (**??**) and δ_{c2} is defined by (B5).

Let δ_{c1} be the value that satisfies

$$k_0^- + k_1^- \cdot \frac{1}{2} = k_0^+ - k_1^+ \cdot \frac{1}{2},$$

i.e.,

$$\begin{aligned}
&c_0 - \delta_{c1} + \frac{c_1}{4N_0}\left(N_0 - 1 + \frac{2}{N_0}\right) + \frac{c_2}{2N_0} + \frac{N_0 - 1}{N_0}\left(\frac{c_1}{2}\frac{N_0 + 2}{N_0} + c_2\right)\frac{1}{2} \\
=&c_0 + \delta_{c1} + \frac{c_1}{4\left(N + N_0\right)}\left(N + N_0 - 1 + \frac{2}{N + N_0}\right) + \frac{c_2}{2\left(N + N_0\right)} - \frac{N + N_0 - 1}{N + N_0}\left(\frac{c_1}{2}\frac{N + N_0 + 2}{N + N_0} + c_2\right)\frac{1}{2}.
\end{aligned}$$

Then when $\delta_c > \delta_{c1}$,

$$\left[k_0^- - k_1^- \cdot \frac{1}{2}, k_0^- + k_1^- \cdot \frac{1}{2}\right] \cup \left[k_0^+ - k_1^+ \cdot \frac{1}{2}, k_0^+ + k_1^+ \cdot \frac{1}{2}\right] = \emptyset.$$

This implies that the under the equilibrium conjectured, when $\tilde{c}_0 = c_0 - \delta_c$, only institutional traders can obtain the order no matter what signals market participants observe.

Let's first verify that it's optimal for any institutional trader to choose $\tilde{s}^-\left(y; \delta_c\right)$ if observing $c_0 + \delta_c$, giving other market participants' strategies. When $\tilde{c}_0 = c_0 - \delta_c$, it's clear that $\tilde{s}^-\left(y; \delta_c\right)$ is an equilibrium if we only have N_0 institutional traders in the market, as suggested by Proposition 1. This is essentially the baseline model of order-by-order auctions with N_0 bidders and unconditional expected inventory cost being $c_0 - \delta_c$. This means that it's optimal for any institutional trader to choose $\tilde{s}^-\left(y; \delta_c\right)$ if there are only $N_0 - 1$ other institutional traders who also choose $\tilde{s}^-\left(y; \delta_c\right)$ and no wholesalers in the market. Adding N wholesalers choosing $\tilde{s}^+\left(y; \delta_c\right)$ does not change this optimality, because given other institutional traders' choice $\tilde{s}^-\left(y; \delta_c\right)$, the N wholesalers will never obtain any order in any state when $\tilde{c}_0 = c_0 - \delta_c$.

Then let's verify that it's optimal for any institutional trader to choose $\tilde{s}^+\left(y; \delta_c\right)$ if observing $c_0 + \delta_c$, given other market participants' strategies. Following our Proposition 1 in the baseline model of order-by-order auctions, $\tilde{s}^+\left(y; \delta_c\right)$ is an equilibrium with $N + N_0$ market makers and unconditional expected inventory cost being $c_0 + \delta_c$. So it's optimal for any institutional trader to choose $\tilde{s}^+\left(y; \delta_c\right)$ if there are other $N + N_0 - 1$ market makers also choosing $\tilde{s}^+\left(y; \delta_c\right)$.

Finally, let's verify that it's optimal for any wholesaler i to choose $\tilde{s}^+\left(y; \delta_c\right)$, given other market

participants' strategies. Suppose the wholesaler i observes a signal y_i, then the wholesaler's utility is

$$U_i = \frac{1}{2}U_1(s_i) + \frac{1}{2}U_2(s_i),$$

where U_1 (U_2) is wholesaler i's profit when the state is $c_0 - \delta_c$ ($c_0 + \delta_c$). It's clear that for any y_i, we have

$$\tilde{s}^+(y_i; \delta_c) = \arg \max_{s_i \in \left(k_0^- + k_1^- \cdot \frac{1}{2}, \infty\right)} U_i,$$

this is because when $s_i \in \left(k_0^- + k_1^- \cdot \frac{1}{2}, \infty\right)$,

$$\left[k_0^- - k_1^- \cdot \frac{1}{2}, k_0^- + k_1^- \cdot \frac{1}{2}\right] \cup \left[k_0^+ - k_1^+ \cdot \frac{1}{2}, k_0^+ + k_1^+ \cdot \frac{1}{2}\right] = \emptyset,$$

and thus we always have

$$U_1(s_i) = 0.$$

And

$$\tilde{s}^+(y_i; \delta_c) = \arg \max_{s_i \in \left(k_0^- + k_1^- \cdot \frac{1}{2}, \infty\right)} U_2.$$

It's also clear that wholesaler will never choose $s_i < k_0^- - k_1^- \cdot \frac{1}{2}$, as any $s_i < k_0^- - k_1^- \cdot \frac{1}{2}$ is dominated by $s_i = k_0^- - k_1^- \cdot \frac{1}{2}$. Suppose that wholesaler choose

$$s_i \in \left[k_0^- - k_1^- \cdot \frac{1}{2}, k_0^- + k_1^- \cdot \frac{1}{2}\right].$$

Note that

$$k_0^- + k_1^- \cdot \frac{1}{2} - \left[\left(c_0 - \underline{\delta_c}\right) - \frac{c_1 + c_2}{2}\right] > 0,$$

and the upper bound of the profit in the case $c_0 - \delta_c$ is

$$k_0^- + k_1^- \cdot \frac{1}{2} - \left[\left(c_0 - \underline{\delta_c}\right) - \frac{c_1 + c_2}{2}\right]$$

because $k_0^- + k_1^- \cdot \frac{1}{2}$ is the highest spread in $s_i \in \left[k_0^- - k_1^- \cdot \frac{1}{2}, k_0^- + k_1^- \cdot \frac{1}{2}\right]$ and $\left(c_0 - \underline{\delta_c}\right) - \frac{c_1+c_2}{2}$ is the lowest inventory cost. Besides, in the case $c_0 + \delta_c$, the wholesaler i will obtain the order with

probability one. And the maximal profit is

$$k_0^- + k_1^- \cdot \frac{1}{2} - \left[\left(c_0 + \underline{\delta_c} \right) - \frac{c_1 + c_2}{2} \right].$$

Then

$$U_1 \leq k_0^- + k_1^- \cdot \frac{1}{2} - \left[\left(c_0 - \underline{\delta_c} \right) - \frac{c_1 + c_2}{2} \right].$$

And

$$U_2 \leq k_0^- + k_1^- \cdot \frac{1}{2} - \left[\left(c_0 + \underline{\delta_c} \right) - \frac{c_1 + c_2}{2} \right].$$

Then

$$U_i \leq \frac{1}{2} U_1 + \frac{1}{2} U_2 \leq k_0^- + k_1^- \cdot \frac{1}{2} - \left[c_0 - \frac{c_1 + c_2}{2} \right].$$

Then

$$U_i < 0 \Longleftarrow k_0^- + k_1^- \cdot \frac{1}{2} - \left[c_0 - \frac{c_1 + c_2}{2} \right] < 0 \Longleftrightarrow \delta_c < \delta_{c2},$$

where

$$\delta_{c2} = \frac{c_1}{4\left(\tilde{N} - N \right)} \left(\left(\tilde{N} - N \right) - 1 + \frac{2}{\tilde{N} - N} \right) + \frac{c_2}{2\left(\tilde{N} - N \right)} + k_1^- \cdot \frac{1}{2} + \frac{c_1 + c_2}{2}. \tag{B5}$$

Then when

$$\delta_c > \underline{\delta} = \max \left\{ \delta_{c1}, \delta_{c2} \right\},$$

we have

$$\tilde{s}^+ \left(y_i; \delta_c \right) = k_0^+ + k_1^+ y_i = \arg \max_{s_i \in (-\infty, \infty)} U_i.$$

This implies that it's optimal for any wholesaler i to choose $\tilde{s}^+ \left(y; \delta_c \right)$, given other market participants' strategies.

Proof of Proposition 8

Since the equilibrium of broker's routing is the same as that in the baseline model, we have

$$\tilde{W}_{total}^{BR} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right),$$

$$\tilde{W}_{I}^{BR} = -\left[c_0 + p_0 \frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)}\right],$$

and

$$\tilde{W}_{W}^{BR} = \frac{p_0\left(2c_1 - p_0 c_1 + Nc_2\right)}{N\left(1+N\right)}.$$

For order-by-order auctions, the welfare outcomes are

$$\begin{aligned}\tilde{W}_{total}^{OBO} =& -\frac{1}{2}\left(c_0 + \delta_c - \frac{N+N_0-1}{N+N_0+1}\frac{c_2}{2}\right) - \frac{1}{2}\left(c_0 - \delta_c - \frac{N_0-1}{N_0+1}\frac{c_2}{2}\right)\\ =& -c_0 + \frac{1}{2}\left(\frac{N+N_0-1}{N+N_0+1} + \frac{N_0-1}{N_0+1}\right)\frac{c_2}{2},\end{aligned}$$

$$\begin{aligned}\tilde{W}_{I}^{OBO} =& -\frac{1}{2}\left[c_0 + \delta_c + \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)}c_1 - \frac{N+N_0-3}{2\left(N+N_0+1\right)}c_2\right]\\ & -\frac{1}{2}\left[c_0 - \delta_c + \frac{1}{N_0\left(N_0+1\right)}c_1 - \frac{N_0-3}{2\left(N_0+1\right)}c_2\right]\\ =& -c_0 - \frac{1}{2}\left[\frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)}c_1 - \frac{N+N_0-3}{2\left(N+N_0+1\right)}c_2 + \frac{1}{N_0\left(N_0+1\right)}c_1 - \frac{N_0-3}{2\left(N_0+1\right)}c_2\right]\end{aligned}$$

and

$$\tilde{W}_{W}^{OBO} = \frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{c_1}{N+N_0} + c_2\right).$$

First,

$$\begin{aligned}&\tilde{W}_{total}^{OBO} > \tilde{W}_{total}^{BR}\\ \Longleftrightarrow& -c_0 + \frac{1}{2}\left(\frac{N+N_0-1}{N+N_0+1} + \frac{N_0-1}{N_0+1}\right)\frac{c_2}{2} > -\left(c_0 - p_0\frac{N-1}{N+1}\frac{c_2}{2}\right)\\ \Longleftrightarrow& \frac{1}{2}\left(\frac{N+N_0-1}{N+N_0+1} + \frac{N_0-1}{N_0+1}\right)\frac{c_2}{2} > p_0\frac{N-1}{N+1}\frac{c_2}{2}.\end{aligned}$$

Then LHS of the above condition is increasing in N_0, let $\underline{N}_0$ be the solution of

$$\frac{1}{2}\left(\frac{N+\underline{N}_0-1}{N+\underline{N}_0+1}+\frac{\underline{N}_0-1}{\underline{N}_0+1}\right)\frac{c_2}{2}=p_0\frac{N-1}{N+1}\frac{c_2}{2}, \tag{B6}$$

then

$$N_0>\underline{N}_0\Longleftrightarrow\tilde{W}_{total}^{OBO}>\tilde{W}_{total}^{BR}.$$

Second,

$$\begin{aligned}
&\tilde{W}_W^{BR}<\tilde{W}_W^{OBO}\\
\Longleftrightarrow&\frac{p_0\left(2c_1-p_0c_1+Nc_2\right)}{N\left(1+N\right)}<\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{c_1}{N+N_0}+c_2\right)\\
\Longleftrightarrow&\frac{p_0\left(2-p_0+N\frac{c_2}{c_1}\right)}{N\left(1+N\right)}<\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{1}{N+N_0}+\frac{c_2}{c_1}\right)\\
\Longleftrightarrow&\frac{p_0\left(2-p_0\right)}{N\left(1+N\right)}+\frac{p_0}{1+N}\frac{c_2}{c_1}<\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{1}{N+N_0}+\frac{c_2}{c_1}\right)\\
\Longleftrightarrow&\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}\right)\frac{c_2}{c_1}>-\left(\frac{1}{2}\frac{N}{\left(N+N_0\right)^2}\frac{1}{N+N_0+1}-\frac{p_0\left(2-p_0\right)}{N\left(1+N\right)}\right)\\
\Longleftrightarrow&\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}\right)\frac{c_2}{c_1}>-\frac{1}{N+N_0}\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{\left(2-p_0\right)\left(N+N_0\right)}{N}\right).
\end{aligned}$$

Since

$$\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}>\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{\left(2-p_0\right)\left(N+N_0\right)}{N},$$

we know that

$$\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}\right)\frac{c_2}{c_1}>-\frac{1}{N+N_0}\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{\left(2-p_0\right)\left(N+N_0\right)}{N}\right)$$

is equivalent to

$$\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}>0\Longleftrightarrow p_0<\frac{1}{2}\frac{N}{N+N_0}\frac{1+N}{N+N_0+1}$$

and

$$\frac{c_2}{c_1} > \frac{-\frac{1}{N+N_0}\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)}{\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}}.$$

Finally,

$$\begin{aligned}
&\tilde{W}_I^{BR} < \tilde{W}_I^{OBO} \\
\Longleftrightarrow &-\left[c_0 + p_0\frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)}\right] \\
&< -c_0 - \frac{1}{2}\left[\frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)}c_1 - \frac{N+N_0-3}{2\left(N+N_0+1\right)}c_2 + \frac{1}{N_0\left(N_0+1\right)}c_1 - \frac{N_0-3}{2\left(N_0+1\right)}c_2\right] \\
\Longleftrightarrow &p_0\frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)} \\
&> \frac{1}{2}\left[\frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)}c_1 - \frac{N+N_0-3}{2\left(N+N_0+1\right)}c_2 + \frac{1}{N_0\left(N_0+1\right)}c_1 - \frac{N_0-3}{2\left(N_0+1\right)}c_2\right] \\
\Longleftrightarrow &p_0\frac{2\left(2-p_0\right) - \left(N-3\right)N\frac{c_2}{c_1}}{N\left(1+N\right)} \\
&> \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)} - \frac{N+N_0-3}{2\left(N+N_0+1\right)}\frac{c_2}{c_1} + \frac{1}{N_0\left(N_0+1\right)} - \frac{N_0-3}{2\left(N_0+1\right)}\frac{c_2}{c_1} \\
\Longleftrightarrow &p_0\frac{2\left(2-p_0\right)}{N\left(1+N\right)} - p_0\left(1-\frac{4}{N+1}\right)\frac{c_2}{c_1} \\
&> \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)} - \frac{1}{2}\left(1-\frac{4}{N+N_0+1}\right)\frac{c_2}{c_1} + \frac{1}{N_0\left(N_0+1\right)} - \frac{1}{2}\left(1-\frac{4}{N_0+1}\right)\frac{c_2}{c_1} \\
\Longleftrightarrow &\left(1-p_0+\frac{4p_0}{N+1} - \frac{2}{N+N_0+1} - \frac{2}{N_0+1}\right)\frac{c_2}{c_1} > \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)} + \frac{1}{N_0\left(N_0+1\right)} - \frac{2p_0\left(2-p_0\right)}{N\left(1+N\right)}.
\end{aligned} \tag{B7}$$

We want to show that if

$$1-p_0+\frac{4p_0}{N+1} - \frac{2}{N+N_0+1} - \frac{2}{N_0+1} \leq 0,$$

we must have

$$\frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)} + \frac{1}{N_0\left(N_0+1\right)} - \frac{2p_0\left(2-p_0\right)}{N\left(1+N\right)} > 0.$$

Since $N > 3$, then both

$$1-p_0+\frac{4p_0}{N+1} - \frac{2}{N+N_0+1} - \frac{2}{N_0+1}$$

and

$$\frac{1}{(N+N_0)(N+N_0+1)}+\frac{1}{N_0(N_0+1)}-\frac{2p_0(2-p_0)}{N(1+N)}>0$$

are decreasing in $p_0 \in (0,1)$. Then it's sufficient to show the above argument holds when $p_0 = 1$, i.e., we need to show that if

$$\frac{1}{N+N_0+1}+\frac{1}{N_0+1}\geq\frac{2}{N+1},$$

we must have

$$\frac{1}{(N+N_0)(N+N_0+1)}+\frac{1}{N_0(N_0+1)}>\frac{2}{N(1+N)}.$$

Note that

$$\begin{aligned}\frac{1}{N+N_0+1}+\frac{1}{N_0+1}\geq\frac{2}{N+1} &\iff \frac{1}{N_0+1}-\frac{1}{N+1}\geq\frac{1}{N+1}-\frac{1}{N+N_0+1}\\ &\iff \frac{N-N_0}{(N_0+1)(N+1)}\geq\frac{N_0}{(N+1)(N+N_0+1)},\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{(N+N_0)(N+N_0+1)}+\frac{1}{N_0(N_0+1)}>\frac{2}{N(1+N)}\\ \iff&\frac{1}{N_0(N_0+1)}-\frac{1}{N(1+N)}>\frac{1}{N(1+N)}-\frac{1}{(N+N_0)(N+N_0+1)}\\ \iff&\frac{1}{N_0(N_0+1)}-\frac{1}{N(N_0+1)}+\frac{1}{N(N_0+1)}-\frac{1}{N(1+N)}\\ &>\frac{1}{N(1+N)}-\frac{1}{N(N+N_0+1)}+\frac{1}{N(N+N_0+1)}-\frac{1}{(N+N_0)(N+N_0+1)}\\ \iff&\frac{N-N_0}{N_0N(N_0+1)}+\frac{1}{N(N_0+1)}-\frac{1}{N(1+N)}>\frac{1}{N(1+N)}-\frac{1}{N(N+N_0+1)}+\frac{N_0}{N(N+N_0)(N+N_0+1)}.\end{aligned}$$

We already know that

$$\frac{1}{N(N_0+1)}-\frac{1}{N(1+N)}\geq\frac{1}{N(1+N)}-\frac{1}{N(N+N_0+1)},$$

then it's sufficient to show

$$\frac{N-N_0}{N_0N(N_0+1)}>\frac{N_0}{N(N+N_0)(N+N_0+1)}\iff\frac{N-N_0}{N_0(N_0+1)}>\frac{N_0}{(N+N_0)(N+N_0+1)}.$$

Since

$$\frac{N-N_0}{(N_0+1)(N+1)} \geq \frac{N_0}{(N+1)(N+N_0+1)},$$

we have

$$\frac{N-N_0}{N_0(N_0+1)} \geq \frac{1}{N_0}\frac{N_0(1+N)}{(N+1)(N+N_0+1)} = \frac{1}{(N+N_0+1)} > \frac{N_0}{(N+N_0)(N+N_0+1)}.$$

Then we show that for the condition (B7), if the LHS

$$1-p_0+\frac{4p_0}{N+1}-\frac{2}{N+N_0+1}-\frac{2}{N_0+1} \leq 0,$$

the RHS

$$\frac{1}{(N+N_0)(N+N_0+1)}+\frac{1}{N_0(N_0+1)}-\frac{2p_0(2-p_0)}{N(1+N)}$$

must be positive. Then the solution to the condition (B7) is

$$1-p_0+\frac{4p_0}{N+1}-\frac{2}{N+N_0+1}-\frac{2}{N_0+1} > 0 \Longleftrightarrow p_0 < \frac{1-\frac{2}{N+N_0+1}-\frac{2}{N_0+1}}{1-\frac{4}{N+1}},$$

and

$$\frac{c_2}{c_1} > \frac{\frac{1}{(N+N_0)(N+N_0+1)}+\frac{1}{N_0(N_0+1)}-\frac{2p_0(2-p_0)}{N(1+N)}}{1-p_0+\frac{4p_0}{N+1}-\frac{2}{N+N_0+1}-\frac{2}{N_0+1}}.$$

Proof of Proposition 9

Consider any $i \in \{1,2...N\}$ and $(x,w) \in \left[-\frac{1}{2},\frac{1}{2}\right] \times \left[-\frac{1}{2},\frac{1}{2}\right]$, let

$$G_0(x|w) = \text{Prob}\left[\min_{-i} w_{-i} \leq x | w_i = w\right]$$

and

$$g_0(x|w) = \frac{dG(x|w)}{dx}.$$

It's easy to show

$$G(x|w) = 1-\left(\frac{1}{2}-x\right)^{N-1},$$

$$g(x|w) = (N-1)\left(\frac{1}{2}-x\right)^{N-2}.$$

Let

$$\begin{aligned}
v(x,w) =& \mathbb{E}\left[\zeta_i | \min_{-i} w_{-i} = x; w_i = w\right] \\
=& p_0 \mathbb{E}\left[\zeta_i | \min_{-i} w_{-i} = x; w_i = w; \exists j \neq i, w_j = y_j = x\right] \\
&+ (1-p_0)\,\mathbb{E}\left[c_i | \min_{-i} w_{-i} = x; w_i = y; \nexists j \neq i, w_j = y_j = x\right] \\
=& p_0 \left[\mathbb{E}(c_0) + \mathbb{E}(c_1)\left(\frac{x + (N-2)\,p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + \mathbb{E}(c_2)\,p_0 w\right] \\
&+ (1-p_0)\left[c_0 + c_1\left(\frac{(N-1)\,p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
=& \bar{c}_0 + \bar{c}_1\left(\frac{\left(p_0 x + (1-p_0)\,p_0 \frac{\frac{1}{2}+x}{2}\right) + (N-2)\,p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + \bar{c}_2 p_0 w.
\end{aligned}$$

We focus on symmetric equilibria. Suppose all of market maker i's opponents use a continuous, increasing bid strategy

$$\bar{B}(w) = \bar{K}_0 + \bar{K}_1 w$$

at time 0. When market maker i observes signal w and reports signal z, its expected profit is

$$\begin{aligned}
U_i(z,w) &= \text{Prob}\left(z \leq \min_{-i} w_{-i} | w_i = w\right)\left[\bar{B}(z) - \mathbb{E}\left(\zeta_i | z \leq \min_{-i} w_{-i}, w_i = w\right)\right] \\
&= [1 - G(z|w)]\left[\bar{B}(z) - \frac{1}{1-G(z|w)}\int_z^1 g(x|w)\,v(x,w)\,dx\right] \\
&= [1 - G(z|w)]\,\bar{B}(z) - \int_z^1 g(x|w)\,v(x,w)\,dx.
\end{aligned}$$

Market maker i's marginal incentive is characterized by

$$\begin{aligned}
&\frac{\partial U_i(z,w)}{\partial z} \\
=&-g(z|w)\bar{B}(z)+(1-G(z|w))\bar{B}'(z)+g(z|w)v(z|w) \\
=&g(z|w)\left[-\bar{B}(z)+\left(\frac{1-G(z|w)}{g(z|w)}\right)\bar{B}'(z)+v(z|w)\right] \\
=&g(z|w)\left[-\bar{K}_0-\bar{K}_1 z+\frac{\frac{1}{2}-z}{N-1}\bar{K}_1+c_0+c_1\left(\frac{\left(p_0 z+(1-p_0)p_0^{\frac{\frac{1}{2}+z}{2}}\right)+(N-2)p_0^{\frac{\frac{1}{2}+z}{2}}+p_0 w}{N}\right)+c_2 p_0 w\right].
\end{aligned}$$

Let's conjecture that in equilibrium we have

$$\left.\frac{\partial U_i(z,w)}{\partial z}\right|_{z=w}=0. \tag{B8}$$

This implies

$$-\left(\bar{K}_0+\bar{K}_1 w\right)+\frac{\frac{1}{2}-w}{N-1}\bar{K}_1+\bar{c}_0+\bar{c}_1\left(\frac{\left(p_0 w+(1-p_0)p_0^{\frac{\frac{1}{2}+w}{2}}\right)+(N-2)p_0^{\frac{\frac{1}{2}+w}{2}}+p_0 w}{N}\right)+\bar{c}_2 p_0 w=0.$$

Since the above condition holds for all w, then $\bar{K}_0$,$\bar{K}_1$ are solved by

$$-\bar{K}_0+\frac{\frac{1}{2}\bar{K}_1}{N-1}+c_0+c_1\frac{N-2}{4N}p_0+\frac{c_1(1-p_0)p_0}{4N}=0$$

$$-\bar{K}_1-\frac{\bar{K}_1}{N-1}+c_2 p_0+\frac{2c_1 p_0}{N}+\frac{c_1(N-2)p_0}{2N}+\frac{c_1(1-p_0)p_0}{2N}=0$$

Then we get

$$\bar{K}_1=\frac{N-1}{N}\left(\bar{c}_2 p_0+\frac{2\bar{c}_1 p_0}{N}+\frac{\bar{c}_1(N-2)p_0}{2N}+\frac{\bar{c}_1(1-p_0)p_0}{2N}\right), \tag{B9}$$

$$K_0=\bar{c}_0+\frac{p_0}{4N^2}\left[\left(3+N^2-p_0-Np_0\right)\bar{c}_1+2N\bar{c}_2\right]. \tag{B10}$$

We also need to verify that condition (B8) is a sufficient condition for optimization. Note that

$g(z|w) > 0$ and

$$-\bar{K}_0 - \bar{K}_1 z + \frac{\frac{1}{2} - z}{N-1}\bar{K}_1 + c_0 + c_1 \left(\frac{\left(p_0 z + (1-p_0)\, p_0 \frac{\frac{1}{2}+z}{2}\right) + (N-2)\, p_0 \frac{\frac{1}{2}+z}{2} + p_0 w}{N} \right) + c_2 p_0 w$$

is linear in z, then it's clear that with (B9) and (B10), we must have that for all w,

$$\frac{\partial U_i(z,w)}{\partial z} < 0 \Longleftrightarrow z > w,$$

confirming that (B8) is a sufficient condition for optimization.

Proof of Lemma 3

Let's introduce the random variable

$$r = \min_i w_i \in \left[-\frac{1}{2}, \frac{1}{2}\right]$$

and its CDF

$$H(r) = 1 - \left(\frac{1}{2} - r\right)^N$$

and PDF

$$h(r) = N\left(\frac{1}{2} - r\right)^{N-1}.$$

First, we know that in our baseline model of broker's routing, the total welfare is

$$W_{total}^{BR} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

The total welfare only depends on inventory allocation but not equilibrium spread, as the equilibrium spread is just a transfer between market makers and investors. In our extension of heterogeneous stocks, it is still the market maker with lowest liquidity signal realization y that obtains the order, so the order allocation is the same as that in our baseline model for any stocks (c_0, c_1, c_2). Then

the total welfare is this extension satisfies

$$W_{heter,total}^{BR} = W_{total}^{BR} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

Since the equilibrium bidding strategy is

$$T(w) = \bar{K}_0 + \bar{K}_1 w,$$

the investor's welfare is

$$\begin{aligned} W_{heter,I}^{BR} &= -\mathbb{E}\left[\bar{K}_0 + \bar{K}_1 r | \min_i w_i = r\right] \\ &= -\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\bar{K}_0 + \bar{K}_1 r\right) N \left(\frac{1}{2} - r\right)^{N-1} dr \\ &= -\left[\bar{c}_0 + p_0 \frac{2(2-p_0)\bar{c}_1 - (N-3)N\bar{c}_2}{2N(1+N)}\right]. \end{aligned}$$

By

$$W_{heter,M}^{BR} = W_{heter,total}^{BR} - W_{heter,I}^{BR},$$

we know

$$\begin{aligned} W_{heter,M}^{BR} &= -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right) + \left[\bar{c}_0 + p_0 \frac{2(2-p_0)\bar{c}_1 - (N-3)N\bar{c}_2}{2N(1+N)}\right] \\ &= (\bar{c}_0 - c_0) + \frac{p_0}{2(N+1)}\left[(N-1)c_2 - (N-3)\bar{c}_2\right] + p_0 \frac{(2-p_0)\bar{c}_1}{N(1+N)}. \end{aligned}$$

Less Is More∗

Bart Zhou Yueshen† Junyuan Zou‡

This version: March 12, 2023

∗ We benefited immensely from discussions with Pierre Collin-Dufresne, Sergei Glebkin, Naveen Gondhi, John Kuong, Benjamin Lester, Joshua Mollner, and Joel Peress. There are no competing financial interests that might be perceived as influencing the analysis, discussion, and/or results of this article.

† INSEAD; b@yueshen.me; 1 Ayer Rajah Avenue, Singapore 138676.

‡ INSEAD; junyuan.zou@insead.edu; Boulevard de Constance, Fontainebleau 77300, France.

Less Is More

Abstract

We show in a model of over-the-counter trading that customers in equilibrium may choose to contact *very few* dealers to incentivize *maximum* liquidity provision—"less is more." This happens when dealers' liquidity supply is sufficiently elastic to competition. This mechanism is orthogonal to conventional concerns, such as contacting or search cost, private information, and relationship. A social planner would mandate even fewer contacts than the market outcome, where customers induce excessive dealer competition. The model predicts endogenous market power, yields implications for regulation and design of electronic platforms, and speaks to customers' search behavior and their execution quality.

Keywords: over-the-counter markets, dealers, trading connections, request-for-quote

(There are no competing financial interests that might be perceived as influencing the analysis, discussion, and/or results of this article.)

1 Introduction

In over-the-counter (OTC) markets, customers approach dealers for their service of liquidity provision. A well-known and robust empirical feature is that customers do *not* reach out to all available dealers. This is true for both conventional phone-based OTC trading and electronic request-for-quote (RFQ) platforms.[1]

At first glance, it might seem beneficial for a liquidity-seeking customer to always contact more dealers: they have larger aggregate capacity to provide more liquidity and are likely to compete more fiercely in price. So what prevents customers from reaching out to all dealers? The literature has pointed to several considerations, for example, search or contact costs, information leakage, and relationship with dealers. (The related literature is reviewed later on p. 5.) This paper turns these channels off and proposes a mechanism that sheds new light on customer-dealer interactions, examines market design implications, and generates testable empirical predictions.

The premise is that it is costly for dealers to provide service (liquidity) to customers. Therefore, dealers strategically choose their service,[2] trading off the marginal service cost and the marginal expected trading gain. One key determinant of a dealer's trading gain is competition—the number of other dealers that the customer is contacting: more competitors, less trading gain, and lower willingness to provide service. A customer thus chooses only a small number of dealers to shield them from too much competition, leaving just enough rent on the table to induce their quality service. In sum,

[1] For example, Hendershott et al. (2020) document that in the corporate bond market, one-third of the customers in their sample contact only one dealer. O'Hara, Wang, and Zhou (2018) show that a customer trades with between one and 19 dealers per bond per year, with at least three-quarters of them trading only with one dealer. In the foreign exchange forward market, Hau et al. (2021) show that an average customer trades only with 1.8 dealers (out of more than 200), and in a later sample, Collin-Dufresne, Hoffmann, and Vogel (2022) find that a customer trades with about three to 13 dealers per month (again, out of more than 200). Evidence specifically regarding RFQ platforms includes: Riggs et al. (2020) report that when trading index credit default swaps (CDS), customers on average query about 4.1 dealers, while the upper bound is 5 on Bloomberg Swap Exchange Facility (SEF) and unrestricted on Tradeweb SEF. Allen and Wittwer (2021) cite annual reports from CanDeal, a multi-dealer platform in Canada, that more than 40% of RFQ auctions did not exhaust the maximum number of dealers allowed.

[2] Dealer service can be thought of as how attentive a dealer is to customers' requests, how much effort they spend in finding inventories for customers, the effectiveness in providing quotes timely and firmly, etc. See, e.g., Bessembinder, Spatt, and Venkataraman (2020) for a review of fixed-income markets and dealers' role and service.

contacting fewer dealers can secure more liquidity provision—"less is more."

Section 2 studies a baseline model, where dealers' service cost is exogenous, to make concrete the above less-is-more mechanism. Section 2.1 sets up the model, and Section 2.2 characterizes the equilibrium. Section 2.3 pinpoints the trade-off that a customer faces: Contacting more dealers positively improves the customer's expected trading gain because there is *better matching*—it is more likely that at least one dealer is able to provide (sufficient) liquidity (timely). However, facing more competition, every dealer expects less trading profit and, consequently, lowers her service to the customer according to the marginal service cost. This novel negative *service effect* hurts the customer, who therefore wants to reduce her dealer contacts.

The analysis further shows that the magnitude of the service effect is governed by dealers' "competition elasticity." In the model, a dealer strategically chooses her service to the customer, wary of how aggressive her competitors are. Intuitively, if more service is provided by others, then less expected trading gain is left, and the dealer reduces her own service by walking down the marginal service cost function. The competition elasticity essentially measures the speed of the "walking down." The larger this elasticity is, the more sensitive are the dealers to each other's service, and the more severe is the negative service effect.

Indeed, in equilibrium, the customer refrains from reaching out to all dealers *only if* the competition elasticity is sufficiently large. Contacting one more dealer is too costly in this case, because the additional competition from this dealer would significantly reduce the customer's overall service from all dealers. Avoiding such a liquidity drought, the customer optimally contacts only few dealers.[3]

Section 3.1 shows that the novel service effect works only if the dealers observe the number of competitors, i.e., the customer's dealer contacts. The reason is that if they do not observe this information, dealers will not be able to react to each others' service competition—the competition elasticity

[3] Although we motivate our model from customer-dealer trades, the less-is-more mechanism can also play a role in inter-dealer trades, and therefore echos the empirical finding that most dealers only trade with very few connected dealers in core–periphery networks. See Maggio, Kermani, and Song (2017), Hollifield, Neklyudov, and Spatt (2017), and Li and Schurhoff (2019) for empirical evidence.

would become zero, thus shutting down the negative service effect. The customer would then see only the matching benefit of more dealers and, contrary to real data, would exhaust all dealers.

Assuming the observability of customers' outside options (the number of the customer's other dealers) is realistic. Private conversations with practitioners suggest that dealers typically know their customers' outside options from, e.g., repeated interactions, due diligence processes, and/or fulfilling compliance requirements. In electronic OTC trading, the number of contacted dealers is directly communicated to the dealers on many RFQ platforms (Riggs et al., 2020). In fact, the model analysis further reveals that customers have an incentive to commit to contacting a subset of dealers.

Section 3.2 studies the regulation and the optimal design of OTC markets. Consider a social planner who can mandate how many dealers a customer should contact. Under mild regularity conditions, the planner always mandates (weakly) fewer dealers than chosen by the customer. In particular, the customer ignores—but the planner accounts for—the intensified dealer competition, which makes the dealers worse off overall. This negative externality concern lends theoretical support for the popular RFQ market design that restricts the maximum number of dealers a customer can contact in each inquiry. Summarizing the above, Section 3.3 makes two specific market design recommendations for RFQ platforms: (i) dealers should always be able to observe how many other dealers a customer is contacting, and (ii) in general it is desirable to constrain customers' dealer contacts, especially if such constraints are made *contingent* on the customer's proposed trade size.

Section 4.1 enriches the baseline model by introducing multiple, possibly heterogeneous, customers and by endowing dealers with certain limited resources (e.g., time, attention, labor, etc.) needed to serve customers. Section 4.2 shows that, in equilibrium, when choosing her service to a particular customer, a dealer trades off the expected trading gain against the *opportunity cost* of spending the limited resources on this customer (as opposed to on other customers). Such an endogenous opportunity cost thus replaces the exogenous service cost in the baseline. In other words, dealers' resource constraint can microfound the premise that dealer service is costly.

Such limited resources are particularly relevant during a short period when, for example, dealers'

infrastructure and hiring are fixed. The model extension, therefore, is well-suited for studying how sudden market stress shocks—such as downgrades of corporate bonds, the volatility in March 2020 due to COVID-19, and the market turmoil caused by UK's "mini-Budget,"—affect customers' behavior in contacting dealers and, in turn, dealers' service to customers. To do so, Section 4.3 considers two groups of customers, non-urgent versus urgent, and examines different forms of stress shocks by varying the total number of customers, the composition of non-urgent and urgent types, and the degree of urgency.

One robust finding is that non-urgent customers always reduce their dealer contacts as the stress shock exacerbates. In fact, it is possible that they completely drop out of trading if the stress becomes severe enough. Intuitively, this is because dealers find it more profitable to allocate their limited resources to serving urgent customers, for they are willing to pay more to trade, and even more so as the market stress shock amplifies their urgency. In other words, non-urgent customers are increasingly "crowded out" by the urgent ones as market stress exacerbates.

Perhaps surprisingly, all customers, not just the non-urgent type, might contact fewer dealers when the market is under stress. This happens when under the stress, more customers become urgent: Facing more urgent customers, dealers understand that their limited resources should earn higher trading gain; that is, each unit of the resource becomes more expensive, bearing a higher opportunity cost. To incentivize dealers to provide such increasingly more expensive service, customers then have to sacrifice further by contacting fewer of them—that is, less is more.

Empirical findings seem to support the above prediction. For example, O'Hara and Zhou (2021) document that when corporate bonds are under fire sell, trading volume via electronic RFQ platforms drops relative to voice trading. That is, consistent with the prediction, when under market stress, customers overall contact fewer dealers by moving away from RFQ platforms, where they simultaneously contact multiple dealers, to conventional voice trading, where it is more difficult and costly to reach multiple dealers.

Contribution and related literature

The paper primarily contributes to the theoretical models that study how customers choose their dealers in OTC trading. The literature has examined several important considerations:

- First, there is exogenous search or contact costs that prevent customers from reaching all dealers. This is seen in early theoretical search models such as Stigler (1961) and applied to OTC markets as in Duffie, Dworczak, and Zhu (2017) and Riggs et al. (2020), among others.
- Second, customers' private information influences how they contact dealers. On the one hand, they may want to use more dealer "connections" to hide their private information, as evidenced by Kondor and Pintér (2022). They may refrain from using too many dealers if the concern of information leakage is dire, as discussed and analyzed by Burdett and O'Hara (1987), Liu, Vogel, and Zhang (2017), Baldauf and Mollner (2022), and Pinter, Wang, and Zou (2022).
- Third, customer-dealer relationship, often modeled in a repeated trading game, can play an important role. For example, Bernhardt et al. (2005) show that relationship endogenously arises and sustains price improvement for the customer, who thus remains with the dealer. Desgranges and Foucault (2005) show that relationship, as in repeated trading, can shield a dealer from being adversely selected by a customer, who, in equilibrium, trades with the dealer only when uninformed. Hendershott et al. (2020) develop a steady-state equilibrium model, where customers choose the number of dealers (i.e., the network size), by trading off the execution speed (the intensity of finding a counterparty) against an exogenous relationship utility flow.

The less-is-more mechanism differs from the above, as there is no exogenous contact cost and no information asymmetry in the one-period trading game.[4]

The paper further contributes to the theory of electronic RFQ platforms. Vogel (2019) studies a hybrid OTC market, with both conventional voice trading and electronic RFQ trading, where both the

[4] Despite the static nature of the model, the less-is-more mechanism helps establish the customer-dealer relationship as well as customers' dealer networks. To see this, one can cast the one-period game in this paper as one in a steady-state equilibrium. The endogenous dealer number, identified by the less-is-more mechanism, then corresponds to the "dealer network size" choice in Hendershott et al. (2020), effectively endogenizing their exogenous relationship utility flows.

dealer number and their response rate (service) are exogenous. In a search setting, Glebkin, Yueshen, and Shen (2022) endogenize dealers' response rate by determining it jointly with the equilibrium asset allocation but keep the number of dealer contacts exogenous. This paper endogenizes both the number of contacts and the response rate. The model suggests novel channels to consider when designing or regulating RFQ platforms, such as whether dealers should be allowed to see how many other dealers customers are contacting, whether an upper bound on the number of contacts should be imposed, etc. Additionally, positive predictions from the model echo existing empirical evidence on RFQ platforms, for example, from Hendershott and Madhavan (2015) and O'Hara and Zhou (2021).

In an independent work, Wang (2022) explores a setting similar to a special case of Levin and Smith (1994) (when the asset value is common knowledge) and finds that customers only want to contact *as few dealers as possible* in RFQ platforms. This is because, in both works, auction bidders (dealers) incur a fixed entry (trading) cost, which implies an infinitely large competition elasticity (as shown in Example 3 in Section 2.3). As a result, the negative service effect becomes extreme, pushing the customer to choose the fewest possible dealers—a corner solution. With a more general service cost function, however, this paper shows that customers' dealer choices can be interior, echoing empirical evidence as seen in, e.g., Riggs et al. (2020) and Allen and Wittwer (2021). Our work thus further contributes to the literature on auctions with endogenous entry (e.g., Levin and Smith, 1994; Menezes and Monteiro, 2000) by highlighting the importance of bidders' competition elasticity.

Existing studies that endogenize dealers' expertise acquisition, such as Glode and Opp (2020) and Li and Song (2021), show that a concentrated market structure (like an OTC market) can incentivize dealers to acquire more expertise to produce valuable information, thus improving social welfare (under certain information structures), compared to a more competitive market structure (like a centralized exchange). Notably, Glode and Opp (2020) share a similar prediction with the less-is-more mechanism that a concentrated OTC market might supply more liquidity to investors than a seemingly more competitive exchange market. Abstracting away from any form of information asymmetry, instead, this paper obtains this result via dealers' costly participation. We explicitly characterize the condition

under which the service effect alone can induce the less-is-more outcome.

The model has additional implications for the execution quality in OTC markets. Following Duffie, Gârleanu, and Pedersen (2005), a large volume of the literature determines the trading price in OTC markets via exogenous Nash bargaining-power parameters. In the current paper, the customer effectively runs a first-price auction among dealers, whose endogenous service in turn determines not only the equilibrium price but also dealers' response rates, trading probability, and trading gain splits—that is, there is *endogenous* bargaining power. The model, therefore, yields rich predictions regarding the execution quality in OTC markets. Notably, O'Hara, Wang, and Zhou (2018) argue that "interacting with a smaller network of dealers can make the [customer] more important to those dealers and hence elicit more favorable executions" (p. 324), and the less-is-more mechanism effectively formalizes this idea. The endogenous dealer response rate and trading probability further speak to Hendershott et al. (2022a), who study the "true cost of immediacy" by accounting also for failed trades.

2 A model of costly dealer service

2.1 Model setup

Agents. There are $\hat{m}$ homogeneous risk-neutral dealers, indexed by $i \in \{1, ..., \hat{m}\}$, where $\hat{m} \geq 2$ is an integer. In this section, we consider one customer, labeling her as customer j (to be consistent with Section 4). We assume that the customer wants to trade one asset. Her trade size is normalized to one unit, and, without loss of generality, we assume that she wants to buy. Her reservation value for the unit is denoted by π_j (> 0), while the dealers value it at 0, thus ensuring positive trading gain.

Timing of events.

1. The customer reaches out to a set $\mathcal{D}_j \subset \{1, ..., \hat{m}\}$ of dealers, with whom she is "in business." Since the dealers are homogeneous, the choice of $\mathcal{D}_j$ simplifies to randomly selecting m_j dealers

out of $\{1, ..., \hat{m}\}$, where $0 \leq m_j \leq \hat{m}$. Below we refer to m_j as the customer's "dealer choice."[5]

2. Every business dealer $i \in \mathcal{D}_j$ observes the customer's type π_j and her dealer choice m_j. Dealer i then privately chooses her "service" for the customer j. We write such service as θ_{ij} with a normalized support of $\in [0, 1]$. Such service is costly: The dealer incurs a cost of $\zeta(\theta_{ij})$ for serving each customer j. We assume that $\zeta(\cdot)$ is convexly increasing, from $\zeta(0) = 0$, and is thrice differentiable, with the first- and the second–order derivatives denoted by $\dot{\zeta}(\cdot)$ and $\ddot{\zeta}(\cdot)$, respectively. We show in Appendix A that assuming a convex $\zeta(\cdot)$ is without loss of generality.
3. Nature makes independent Bernoulli draws $\{A_{ij}\}_{i \in \mathcal{D}_j}$ with respective success rates $\{\theta_{ij}\}_{i \in \mathcal{D}_j}$. We say a dealer i is "ready" for the customer j if $A_{ij} = 1$. Only when ready can a dealer i respond to the customer j, by making a take-it-or-leave-it offer (TIOLIO) at price p_{ij}. No dealer observes whether others are ready.
4. The customer j then compares all available TIOLIOs and chooses the best price p_j, i.e.,

$$p_j = \arg\min_{p \in \{p_{ij} \,|\, A_{ij}=1\}} p \tag{1}$$

to trade with the quoting dealer. If there are multiple dealers quoting the same best price, the customer randomly chooses one to trade with. If there is no offer, there is no trade.

Equilibrium. The equilibrium is characterized by three sets of endogenous objects: (i) the customer's dealer choice m_j; (ii) the dealers' service $\{\theta_{ij}\}$; and (iii) the dealers' quotes p_{ij} (when $A_{ij} = 1$). All agents maximize their respective expected trading profits. The analysis below focuses on symmetric equilibria in which the homogeneous dealers choose the same (ii) and (iii).

Remarks

Remark 1 (Customer's reservation value). By normalizing the homogeneous dealers' reservation value to zero, the customer's reservation value π_j is the expected gains from trade. Such trading gains can arise from, for example, the customer's urgency to trade (willingness to trade), hedging need, and

[5] We assume away costs associated with the dealer choice. This differentiates our model from, e.g., Riggs et al. (2020).

sentiment.

Remark 2 (Dealers' learning about clients). We assume that each dealer $i \in \mathcal{D}_j$ perfectly observes both π_j and m_j of the customer j, because of the non-anonymity of OTC markets. For example, a dealer needs to do her due diligence, e.g., to "know your customers (KYC)." Alternatively, dealers can also learn about $\{\pi_j, m_j\}$ from repeated interactions (which we do not explicitly model) with the customer. The assumption that dealers can *perfectly* observe $\{\pi_j, m_j\}$ is not as restrictive as meets the eye: For π_j, what matters is the *expected* gains from trade, and we only need to assume such an expectation exists. For m_j, as will be shown in Section 3.1, the customer, in fact, has incentive to truthfully reveal this information to her dealers (and commit to it).

Remark 3 (Dealers' costly service and readiness). Dealers serve their customers by providing timely trading opportunities, for example, by arranging the inventory that the customer wants (or providing inventory space when the customer seeks to sell). We model the quality of such service via $\theta_{ij} \in [0, 1]$, a higher value of which indicates, for example, more effort by the dealer to arrange the inventory wanted. Only when the inventory is successfully arranged (i.e., when $A_{ij} = 1$) is the dealer "ready" to quote to the customer. Thus θ_{ij} also reflects the timeliness and the firmness of a dealer's quote. Such effort to arrange inventory is costly. There is labor costs, like hiring professionals to cover trading desks day and night, doing risk management and due diligence, and fulfilling regulatory compliance requirements. In addition, serving timely and firm quotes means commitments to trade, implying costly margins and collaterals for arranging inventories and for clearing. These service costs are summarized in $\zeta(\theta_{ij})$, which we later endogenize in Section 4 via dealers' resource constraints. Since such service or effort is a dealer's hidden action, we assume that θ_{ij} is unobservable by other dealers.

Remark 4 (RFQ trading). Our setup closely matches many electronic trading platforms that adopt the RFQ protocol. In such platforms, a customer endogenously chooses m_j, the number of dealers from whom she requests a quote. In doing so, the customer's intended trade size and side, as well as her identity, are revealed to the dealers (Riggs et al., 2020; O'Hara and Zhou, 2021), who can thus observe (or estimate) the trading gain π_j. However, depending on the platform, m_j may or may not be

observed by dealers. For example, "dealers observe how many other dealers a customer contracts" on Bloomberg SEF and Tradeweb SEF (p. 858, Riggs et al., 2020); but on MarketAxess, "[t]he dealers do not know the number or identities of the other dealers contacted" (p. 370, O'Hara and Zhou, 2021). We discuss this contrast in the market design through the lens of a welfare analysis in Sections 3.1–3.2.

Remark 5 (Voice trading). Our setup also speaks to conventional voice trading in OTC markets. Such voice trading is typically modeled as bilateral meetings between a customer and a dealer (when they are matched), as in, e.g., Duffie, Gârleanu, and Pedersen (2005). We argue that a customer can instead approach multiple dealers, especially when she seeks to execute a trade in a timely fashion. A case in point is the Public Sector Purchase Program (PSPP) by European Central Bank: when purchasing a bond, the executing central bank approaches multiple dealers to seek quotes and then trades at the best price (Hammermann et al., 2019), effectively running a first-price auction among selected dealers as in our model. Breckenfelder, Collin-Dufresne, and Corradin (2022) study the PSPP via a similar first-price auction model.

2.2 Equilibrium analysis

We analyze the equilibrium backwards. Section 2.2.1 first solves dealers' quoting strategy $\{p_{ij}\}$ (if they are ready to quote), assuming symmetric service to the same customer j; that is, $\theta_{ij} = \theta_j$ for all $i \in \mathcal{D}_j$. Section 2.2.2 then looks for a Nash equilibrium, where the symmetric service θ_j is a function of dealers' information $\{m_j, \pi_j\}$ about the customer j. Finally, Section 2.2.3 studies the customer j's optimal dealer choice m_j.

2.2.1 Dealers' quoting

Consider a dealer $i \in \mathcal{D}_j$, i.e., a business dealer of the customer j, who is ready to quote ($A_{ij} = 1$). The dealer would like to capture the full surplus by quoting $p_{ij} \uparrow \pi_j$, just below the customer's reservation value. However, she faces $(m_j - 1)$ *potential* competitors, as their quotes (ask prices) might be lower

than p_{ij}. Yet, each competitor $i' \in \mathcal{D}_j$ (and $i' \neq i$) is able to quote only probabilistically (when $A_{i'j} = 1$). That is, the dealers in $\mathcal{D}_j$ engage in a price competition against *unknown number of competitors*.

Such price competition differs from the standard Bertrand competition, in which every dealer quotes her reservation price of $p_{ij} = 0$ and the customer gets the full surplus π_j. Here, every dealer $i \in \mathcal{D}_j$ has the incentive to charge a higher price, $p_{ij} = \alpha_{ij}\pi_j$ for some $\alpha_{ij} \in (0, 1]$. This is because she might actually be the only dealer who is ready, in which case her TIOLIO at p_{ij} is the only available offer to the customer. As long as $\alpha_{ij} \leq 1$, the customer j will accept it and the dealer i pockets the profit of $p_{ij} = \alpha_{ij}\pi_j$. In a Nash equilibrium, however, the fraction α_{ij} cannot be deterministic, as the undercutting argument of Bertrand competition will drive $\alpha_{ij} \downarrow 0$, and yet, in this case, it would be strictly better off to quote some $\alpha_{ij} > 0$. This heuristic discussion is formalized in the proof and summarized by the following lemma.

Lemma 1 (Mixed-strategy quoting). Suppose the dealers in $\mathcal{D}_j$ have followed a symmetric strategy to provide service $\theta_{ij} = \theta_j$ (> 0) to the customer j. Then there exists a unique mixed-strategy equilibrium, in which each dealer i with $A_{ij} = 1$ quotes $p_{ij} = \alpha_{ij}\pi_j$, where α_{ij} is a random variable, i.i.d. across i, with c.d.f. $F(\alpha_{ij}; \theta_j, m_j) := \frac{1}{\theta_j} - \left(\frac{1}{\theta_j} - 1\right)\alpha_{ij}^{-\frac{1}{m_j - 1}}$, distributed on $\alpha \in \left[(1-\theta_j)^{m_j-1}, 1\right]$.

Note that when $m_j = 1$, $F(\cdot)$ degenerates to a single probability mass at the maximum $\alpha_{ij} = 1$. We can then use the above lemma to compute dealer and customer's respective expected trading gains.

Lemma 2 (Endogenous split of trading gain). Under Lemma 1, a dealer i who is ready to quote $(A_{ij} = 1)$ expects a revenue of

$$\left(1 - \theta_j\right)^{m_j - 1}\pi_j \tag{2}$$

when quoting to the customer j. Furthermore, the customer j expects a trading gain of

$$\pi_j^{\mathrm{c}} := \left(1 - (1-\theta_j)^{m_j} - m_j\theta_j \cdot (1-\theta_j)^{m_j - 1}\right)\pi_j. \tag{3}$$

These expressions can be interpreted as follows. Under the mixed strategy given in Lemma 1, a dealer who is ready to quote $(A_{ij} = 1)$ must be indifferent from choosing any price in the relevant support.

In particular, if she chooses $p_{ij} \uparrow \pi_j$, then she wins the price competition, earning π_j, only if all her competitors are absent, which happens with probability $(1-\theta_j)^{m_j-1}$. Note that unconditionally, the dealer therefore expects $\theta_{ij} \cdot (1-\theta_j)^{m_j-1}\pi_j$, which is monotone increasing in the dealer's service θ_{ij}. This is consistent with the evidence from Hendershott et al. (2022b) that more active dealers have more order flow. In Section 2.2.2, we use this expression to derive dealers' optimal service choice θ_j.

As given in (3), the customer j expects a fraction of the total trading gain π_j. This fraction is less than 1, for two reasons: (i) with probability $(1-\theta_j)^{m_j}$, none of her m_j dealers is ready and there is no trade; and (ii) each of the m_j dealers is ready with probability θ_j and, in that case, expects (2). This fraction is strictly positive, implying that even though the customer only faces TIOLIOs, she has endogenous bargaining power, due to the above price competition among dealers. Section 2.2.3 uses (3) to derive the customer's optimal dealer choice m_j.

2.2.2 Dealers' service to the customer

Consider a dealer $i \in \mathcal{D}_j$. She knows that the number of competing dealers is $m_j - 1$. She also takes as given these competing dealers' symmetric service choice of $\theta_{i'j} = \theta_j$, $\forall i' \in \mathcal{D}_j$ and $i' \neq i$. Using (2), before A_{ij} realizes, dealer i expects a payoff of $\theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1}\pi_j$, where θ_{ij} is her service to client j. In Appendix A, we show that it suffices to consider only a pure strategy of θ_{ij}, thanks to the convexity of the service cost $\zeta(\cdot)$. Therefore, dealer i's problem is

$$\max_{\theta_{ij}\in[0,1]} \theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1}\pi_j - \zeta\left(\theta_{ij}\right) \tag{4}$$

Its solution is characterized by the following proposition.

Proposition 1 (Dealers' symmetric service). In a symmetric-strategy equilibrium, every dealer $i \in \mathcal{D}_j$ chooses the same service $\theta_{ij} = \theta_j$ for customer j:

$$\theta_j = \mathbb{1}_{\{\pi_j > \dot{\zeta}(0)\}} g\left(m_j, \pi_j\right),$$

where $g(\cdot)$ is an implicit function of θ_j, given by $\left(1-\theta_j\right)^{m_j-1}\pi_j = \dot{\zeta}(\theta_j)$, and $\mathbb{1}_{\{\cdot\}}$ is an indicator.

We provide some intuition here and leave the formal proof to the appendix. The implicit function $g(\cdot)$ is defined by the first-order condition of (4) with respect to θ_{ij}:

$$\left(1-\theta_j\right)^{m_j-1}\pi_j - \dot{\zeta}(\theta_j) = 0, \tag{5}$$

with the symmetric $\theta_{ij} = \theta_j$. In words, $g(\cdot)$ solves the symmetric θ_j that equates the marginal benefit and cost: The marginally higher probability to win the price competition and earn (2) must break even with the marginal cost of $\dot{\zeta}(\theta_j)$. The solution $g(\cdot)$, however, might be constrained by the requirement of $\theta_j \in [0,1]$. In particular, we show in the proof that only the lower bound $\theta_j \geq 0$ might bind, hence the indicator function in the proposition.

An important implication of Proposition 1 is that the optimal service θ_j decreases in m_j. Assuming m_j as a nonnegative real number,[6] then the following derivative is well-defined:

$$\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} = \mathbb{1}_{\{\pi_j > \dot{\zeta}(0)\}} \cdot \frac{(1-\theta_j)\ln(1-\theta_j)}{m_j - 1 + (1-\theta_j)\ddot{\zeta}(\theta_j)/\dot{\zeta}(\theta_j)} \leq 0. \tag{6}$$

Indeed, if there are too many potential competitors (large m_j), doing business with the customer is not going to be very profitable, and there is no point providing much costly service to her.

2.2.3 The customer's choice of dealers

The customer j, before trading starts, chooses m_j dealers to maximize her ex-ante expected trading gain π_j^{c}, given by (3), subject to dealers' optimal service θ_j (Proposition 1).

Proposition 2 (Customers' dealer choice). If $\pi_j \leq \dot{\zeta}(0)$, the customer will not trade and is indifferent to choosing any $m_j \in [0, \hat{m}]$. If $\pi_j > \dot{\zeta}(0)$, there always exists some $m_j \in (1, \hat{m}]$ that maximizes the customer's ex-ante payoff π_j^{c}, as given in (3).

Several features of the proposition are worth highlighting. First, only if the customer is "large" enough will she approach dealers initially—that is, if the trading gain π_j is too small ($\leq \dot{\zeta}(0)$), no dealer will

[6] While it is natural to think of m_j as an integer (number of dealers), for simplicity, we treat it as a nonnegative real number. That is, we allow the customer to contact, for example, $m_j = 4.7$ dealers, with the rough interpretation that she plays a mixed strategy between choosing 4 and 5 dealers.

serve her ($\theta_j = 0$, Proposition 1) and, knowing this, this customer j would not bother to open accounts with dealers. (More precisely, she is indifferent to contacting any dealer or not, as none will serve her.)

Second, there is a lower bound of $m_j > 1$ (if $\pi_j > \dot{\zeta}(0)$). Intuitively, doing business with only one dealer effectively waives the dealer from competition with others. As such, the dealer extracts all the trading gain, and the customer expects $\pi_j^c = 0$, following (3) with $m_j = 1$. Instead, choosing any $m_j > 1$ induces at least some competition among dealers, capturing some $\pi_j^c > 0$.

Third, the proposition is only about the existence of the optimal m_j. Such existence readily follows the fact that the support of m_j is bounded by $[0, \hat{m}]$ and that the objective π_j^c is continuous in m_j. We provide a more detailed characterization of the optimal m_j in Section 2.3, where we discuss when m_j is interior or cornered and when it is unique.

2.2.4 Summary of equilibrium

In summary, the equilibrium is as follows:

(i) The customer j chooses m_j dealers for her $\mathcal{D}_j$, where m_j is given in Proposition 2.

(ii) Every dealer $i \in \mathcal{D}_j$ provides symmetric service $\theta_{ij} = \theta_j$ as given in Proposition 1.

(iii) If $A_{ij} = 1$, then dealer $i \in \mathcal{D}_j$ quotes an ask price p_{ij} according to Lemma 1.

For the subsequent analysis to be meaningful, we focus on the case of a large customer with $\pi_j > \dot{\zeta}(0)$ in the rest of Section 2, for otherwise there is no trading (Proposition 2).

2.3 When less is more

One key result of our model is that customers do not always exhaust the available dealers; that is, they do business with fewer dealers to maximize their expected trading gains—less is more. Mathematically, this requires the optimal dealer choice m_j to be interior, $1 < m_j < \hat{m}$. To study when this happens, we decompose the effects of a marginally larger m_j on the customer's expected trading gain π_j^c by examining the derivative of π_j^c with respect to m_j.

Lemma 3 (Customer's tradeoff). Suppose π_j^{c}, as given by (3), is differentiable in m_j. Then

$$\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = \underbrace{\frac{\partial \pi_j^{\mathrm{c}}}{\partial m_j}}_{\geq 0,\text{ direct effect}} + \overbrace{\frac{\partial \pi_j^{\mathrm{c}}}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}}^{\leq 0,\text{ indirect effect}},$$

where the direct effect is always positive and the indirect effect is always negative.

That is, by chain rule, we see a pair of opposing effects:

- **Matching effect:** A larger m_j helps the customer reach more dealers, who will more likely be able to serve her when she needs to trade and will compete more fiercely to provide better quotes. This is the direct effect of $\frac{\partial \pi_j^{\mathrm{c}}}{\partial m_j}$, and it is always positive, inducing the customer to contact as many dealers as possible.
- **Service effect:** On the other hand, as m_j increases, dealers know that they face more competition and expect less revenue. Hence, the lowered expected revenue drives them to *reduce* their service to the customer. This novel indirect service effect is always negative, because dealers reduce their service facing more competition, following (6).

A key determinant in the net effect of $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j}$ is dealers' "competition elasticity," defined as

$$\varepsilon := \frac{\mathrm{d}\left(\ln\left(1-\theta_{ij}\right)\right)}{-\mathrm{d}\left(\ln\left(1-\theta_j\right)^{m_j-1}\right)}. \tag{7}$$

In words, ε (> 0) captures how sensitive a dealer i is to competition: If the competing dealers serve more to customer j (reducing their no-service probability by $\mathrm{d}\left(\ln\left(1-\theta_j\right)^{m_j-1}\right)$), dealer i will serve less (increasing her own no-service probability by $\mathrm{d}\left(\ln\left(1-\theta_{ij}\right)\right)$). The larger (more positive) ε is, the more aggressively dealer i reduces her service. In other words, the competition elasticity (7) effectively measures the strength of the service effect. If ε is sufficiently large, the service effect dominates, thus making the customer unwilling to reach out to more dealers.

Under the optimal symmetric service θ_j given by Proposition 1, the competition elasticity can be simplified. In particular, for $\pi_j > \dot{\zeta}(0)$, dealer i's first-order condition (5) holds with $\theta_{ij} = \theta_j > 0$.

Substituting the (5)-implied $\left(1-\theta_j\right)^{m_j-1} = \dot{\zeta}(\theta_j)/\pi_j$ into the denominator of (7), we obtain

$$\varepsilon(\theta_j) = \frac{1}{1-\theta_j}\frac{\dot{\zeta}(\theta_j)}{\ddot{\zeta}(\theta_j)}, \text{ for } \theta_j \in (0,1). \tag{8}$$

That is, given the dealers' optimal service (Proposition 1), the competition elasticity depends only on the shape of the service cost $\zeta(\cdot)$. Below we study $\varepsilon(\cdot)$ to characterize when the customer's equilibrium choice of m_j is interior and when it is unique.

2.3.1 Interior solution with sufficiently many dealers

To examine when m_j is interior, we first relax the customer's dealer choice from $m_j \in [1, \hat{m}]$ to $m_j \in [1, \infty)$. This avoids the "mechanical" corner solution when $\hat{m}$ is too small—for example, if $\hat{m} = 1$, then a corner solution of $m_j = \hat{m} = 1$ always arises. A sufficient condition for $m_j < \infty$ is given below.

Proposition 3 (Not using infinitely many dealers m_j). When there are sufficiently many dealers ($\hat{m} \to \infty$), the customer j's optimal dealer choice m_j is finite if

$$\varepsilon(0) > 2. \tag{9}$$

Furthermore, $\varepsilon(\theta_j) > 2$ at this optimal m_j, where θ_j is the optimal dealer service given by Proposition 1.

Intuitively, condition (9) effectively requires the competition elasticity ε to be sufficiently large, so that the negative service effect is severe enough to deter the customer from reaching out to too many dealers.

Below we introduce a general class of service cost functions, which can ensure sufficiently large $\varepsilon(0)$ as required by (9):

$$\zeta(\theta) = \begin{cases} \frac{a}{1-b}\left(1-(1-\theta)^{1-b}\right) + c\theta, \text{ if } b \neq 1; \text{ and} \\ -a\ln(1-\theta) + c\theta, \text{ if } b = 1. \end{cases} \tag{10}$$

The competition elasticity, under this class of $\zeta(\cdot)$, can be found as $\varepsilon(0) = \frac{a+c}{ab}$, and it satisfies (9) for various parameter values of $\{a, b, c\}$. In particular, (10) nests the following special cases.

Example 1 (Constant competition elasticity)**.** If the parameters satisfy $a > 0$, $b > 0$, and $c = 0$, then this class of $\zeta(\cdot)$ is convexly increasing and satisfies $\zeta(0) = 0$, as assumed in Section 2.1. Furthermore, the competition elasticity is constant, $\varepsilon(\theta) = 1/b$, satisfying (9) if $b < \frac{1}{2}$. Such a cost function $\zeta(\cdot)$ is reminiscent of constant relative risk aversion (CRRA) utility functions.

Example 2 (Linearly decreasing competition elasticity)**.** If $a > 0$, $b = 1$, and $c > 0$, it can be seen that the resulting $\zeta(\cdot)$ is also convexly increasing and satisfies $\zeta(0) = 0$, as assumed in Section 2.1. The competition elasticity becomes $\varepsilon(\theta) = 1 + \frac{c}{a}(1 - \theta)$ and satisfies (9) if $c > a$.

Example 3 (Infinitely large competition elasticity)**.** If $a = 0$, the cost function becomes $\zeta(\theta) = c\theta$. This linear service cost can be seen as the result of dealers paying a fixed cost of $c > 0$ only when ready ($A_i = 1$), for example, due to regulatory compliance, clearing requirements, or risk management. Jovanovic and Menkveld (2022) assume such a cost function to study quote dispersion in limit order markets. In particular, the constant competition elasticity becomes $\varepsilon \uparrow \infty$, satisfying (9). Wang (2022) also assumes such a cost function and, because of the infinite competition elasticity, finds that the customer only wants to contact as few dealers as possible.

Example 4 (Quadratic service cost)**.** If $b = -1$, then the cost function becomes $\zeta(\theta) = -\frac{a}{2}\theta^2 + (a+c)\theta$. It is also convexly increasing and satisfies $\zeta(0) = 0$, as assumed in Section 2.1, if $a < 0$ and $a + c > 0$. The implied competition elasticity is $\varepsilon(\theta) = -\frac{c}{a}\frac{1}{1-\theta} - 1$ and satisfies (9) if $c > -3a$.

2.3.2 Interior solution with finite dealers

We now return to the more realistic setting of finite dealers, i.e., $\hat{m} < \infty$. To facilitate subsequent analyses, we impose a regularity condition to ensure that π^{c} is quasi-concave in m_j.

Lemma 4 (A sufficient condition for uniqueness). It is sufficient to assume that

$$\frac{\mathrm{d}\varepsilon(\theta)}{\mathrm{d}\theta} \leq 0 \text{ for all } \theta \in [0, 1] \tag{11}$$

to ensure that π_j^{c} is quasi-concave on $m_j \in (1, \infty)$.

To see the intuition, note that the benefit of increasing m_j—the positive matching effect—always diminishes with m_j.[7] On the cost side, the service loss exacerbates with m_j. This is because when m_j is small (large), each dealer knows that she faces low (high) competition and will provide a lot of (little) service to the customer, i.e., θ_j is high (low). The monotone decreasing $\varepsilon(\cdot)$ then implies that an increase in m_j reduces a large (small) amount of service θ_j when m_j is large (small). In other words, the negative service effect, following $\frac{\mathrm{d}\varepsilon}{\mathrm{d}\theta} \leq 0$, is more severe when m_j is large—the customer's cost of losing service exacerbates with m_j. Combining the diminishing benefit and the exacerbating cost, the quasi-concavity guarantees that the optimal m_j is unique.

It is worth emphasizing that the condition (11) is sufficient but *not necessary* for the optimal m_j to be unique in the support of $[1, \hat{m}]$. In Example 4, for instance, $\varepsilon(\theta)$ is monotone increasing in θ, thus not satisfying (11), but it can still be shown that the customer's objective π_j^{c} remains quasi-concave. (Examples 1–3 clearly satisfy (11).)

With the help of (9) and (11), we can now obtain additional useful comparative statics and, further, refine the equilibrium characterization of m_j given earlier in Proposition 2.

Corollary 1 (When less is more). Assume (9) and (11). Then the customer j's optimal dealer choice m_j is unique. Further, both m_j and the dealers' optimal service θ_j are (weakly) increasing in π_j. In particular, the customer chooses fewer dealers than available, i.e., $m_j < \hat{m}$, if and only if

$$\pi_j < \frac{\dot{\zeta}(\hat{\theta})}{(1-\hat{\theta})^{\hat{m}-1}}, \tag{12}$$

where $\hat{\theta} \in (0,1)$ is a unique exogenous threshold given by (B.4) in the proof.

Intuitively, dealers compete more fiercely for larger customers; that is, all else being equal, a customer with larger π_j receives more service θ_j. Hence, increasing m_j induces more service from all dealers for a customer with larger π_j. Further, under (11), the competition elasticity ε is (weakly) lower with more service θ_j, thus weakening the negative service effect of increasing m_j. Therefore, both of these

[7] Recall from (6) that m_j and θ_j are negatively related. Therefore, when m_j is small, θ_j is large, and a marginal increase in m_j increases the trading probability significantly by such a large θ_j. If m_j has become very large, each of the dealers provides very low θ_j, as does the marginal additional dealer, adding very little to the trading probability.

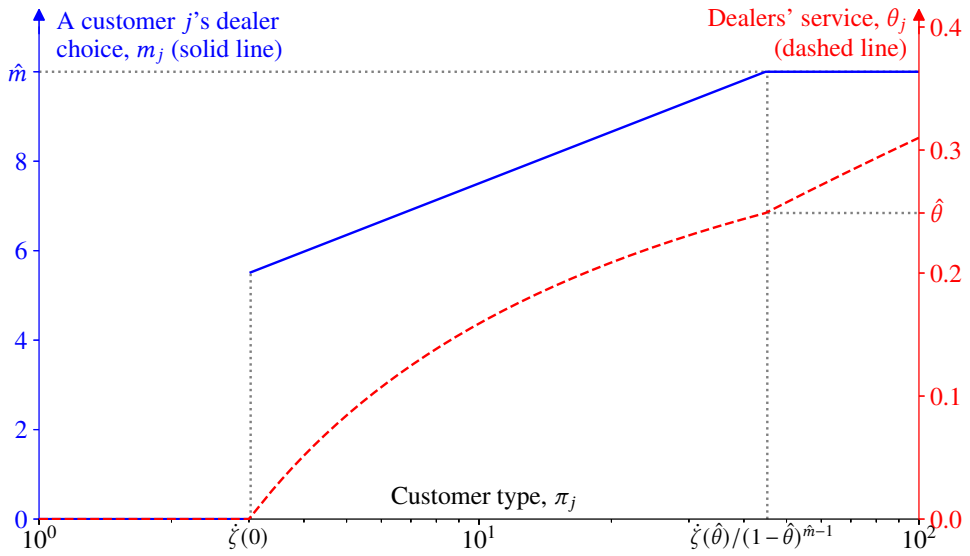


Figure 1: A customer's dealer choice and dealers' service to her. This figure plots the equilibrium m_j (solid line, left axis) and θ_j (dashed line, right axis) as functions of customer type π_j, varying in $\pi_j \in [10^0, 10^2]$ on the horizontal axis (log scale). Dealers' service cost function $\zeta(\cdot)$ is parameterized as in Example 1, with $a = 3.0$ and $b = 0.44$. The total number of dealers is set at $\hat{m} = 10$.

effects incentivize a larger customer (larger π_j) to reach out to more dealers.[8]

Figure 1 illustrates the patterns. The customer's π_j is plotted on the horizontal axis in log scale. The solid line (left axis) shows that she only reaches out to $m_j > 0$ dealers if she is "large enough," i.e., when $\pi_j > \dot{\zeta}(0)$. As π_j increases, she does business with more dealers, until she exhausts all of them, i.e., $m_j = \hat{m}$ for $\pi_j > \dot{\zeta}(\hat{\theta})/(1-\hat{\theta})^{\hat{m}-1}$. Dealers' symmetric service θ_j, the dashed line (right axis), also increases with π_j as, intuitively, dealers are willing to provide more service for larger customers. Notably, however, its initially increase is slower than later, when m_j is capped at $\hat{m}$. This is because initially there are new, competing dealers introduced by the customer's increasing m_j, and such competition on the extensive margin dampens the existing dealers' incentive to serve the customer. Once $m_j = \hat{m}$ is capped, such an extensive-margin competition stops, allowing θ_j to increase faster with π_j.

[8] Note also that (12) can be equivalently rewritten as $\hat{m} > 1 + \frac{\ln(\dot{\zeta}(\hat{\theta})/\pi_j)}{\ln(1-\hat{\theta})}$. That is, it is essentially a variation of "sufficiently many dealers" as studied in Section 2.3.1. In other words, because of (11), the requirement of a sufficiently large $\hat{m}$ can be translated into a requirement of small π_j to ensure interior m_j.

3 Market design implications

A key assumption in the model is that every dealer of a customer i observes the customer's dealer choice m_j. In conventional OTC trading, such observability can arise from dealers' due diligence exercises and/or repeated interactions with the customer. On RFQ platforms, such observability is a market design choice—indeed, some, but not all, RFQ platforms reveal m_j to dealers (see Remark 4). This section studies related market design issues for RFQ platforms. To set the stage, Section 3.1 first studies the observability of m_j. Section 3.2 then examines welfare implications. Finally, Section 3.3 makes concrete market design suggestions.

3.1 The observability of the customer's dealer choice

To illustrate the idea, this subsection considers an RFQ platform where the customer can choose, before trading starts, whether to reveal her choice m_j to her dealers. If she chooses to reveal so, the equilibrium characterized in Sections 2.2–2.3 applies.

What will happen if she does not reveal m_j? The customer still chooses $m_j \in [0, \hat{m}]$ to maximize her expected payoff π_j^c as given by (3). However, her dealers' (symmetric) service θ_j can no longer be a function of the unobservable m_j. Assuming differentiability, therefore,

$$\frac{\mathrm{d}\pi_j^c}{\mathrm{d}m_j} = \underbrace{\frac{\partial \pi_j^c}{\partial m_j}}_{\geq 0,\ \text{direct effect}} + \overbrace{\frac{\partial \pi_j^c}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}}^{=0,\ \text{indirect effect}} = \frac{\partial \pi_j^c}{\partial m_j} \geq 0.$$

Compared to the decomposition in Lemma 3, it can be seen that the indirect negative service effect, which used to balance the direct positive matching effect, is no longer in effect, because $\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} = 0$. With $\frac{\mathrm{d}\pi_j^c}{\mathrm{d}m_j} > 0$, the customer j will then contact as many dealers as possible, i.e., $m_j = \hat{m}$.

Consequently, the dealers in equilibrium also know that $m_j = \hat{m}$. Their symmetric optimal service choice θ_j is then a special case of Proposition 1, with $m_j = \hat{m}$. Recall from (6) that $\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} \leq 0$. Therefore,

the customer, in fact, gets the lowest service of

$$\underline{\theta}_j := \mathbb{1}_{\{\pi_j > \dot{\zeta}(0)\}} g(\hat{m}, \pi_j).$$

Intuitively, this is because the customer always contacts all dealers, intensifying their competition and driving down their profit, which no longer justifies any quality service. In turn, this lowest service $\underline{\theta}_j$ makes the customer (weakly) worse off.

Proposition 4 (Truthful revelation of m_j). Assume (9) and (11). Every customer j individually (weakly) prefers truthfully revealing her dealer choice m_j. That is, $\pi_j^{\mathrm{c}}(m_j) \geq \pi_j^{\mathrm{c}}(\hat{m})$, where m_j is the equilibrium outcome given in Corollary 1; and the inequality is strict if $\dot{\zeta}(0) < \pi_j < \dot{\zeta}(\hat{\theta})/(1-\hat{\theta})^{\hat{m}-1}$.

Proposition 4 also shows that a sufficiently large customer ($\pi_j \geq \dot{\zeta}(\hat{\theta})/(1-\hat{\theta})^{\hat{m}-1}$) is indifferent about revealing her m_j or not. This is because the dealers are okay with not observing her m_j, as they know that such a large customer does business with all dealers no matter what.

3.2 Welfare and customers' dealer choice

The above analysis shows that customers weakly prefer that the RFQ platform directly reveals their dealer choices m_j to the contacted dealers. Do dealers also benefit from the observability of m_j? Is trading more efficient overall? In this subsection, we study how welfare is affected by m_j, before continuing with concrete market design suggestions in Section 3.3.

A general expression of welfare. Suppose the customer j contacts m_j dealers and receives an amount of θ_{ij} service from dealer $i \in \mathcal{D}_j$. The trading gain of π_j is realized as long as at least one dealer out of m_j is ready, i.e., with probability $1 - \prod_{i\in\mathcal{D}_j}(1-\theta_{ij})$. To provide such service, a dealer i incurs a cost of $\zeta(\theta_{ij})$. Therefore, welfare is calculated as

$$w = \left(1 - \prod_{i\in\mathcal{D}_j}(1-\theta_{ij})\right)\pi_j - \sum_{i\in\mathcal{D}_j}\zeta(\theta_{ij}). \tag{13}$$

For example, if dealer service is symmetric, $\theta_{ij} = \theta_j$, then welfare becomes

$$w = \left(1 - (1 - \theta_j)^{m_j}\right)\pi_j - m_j\zeta(\theta_j). \tag{14}$$

Social planner mandating m_j. Below we study how a social planner mandates the customer's dealer choice m_j to maximize welfare and compare the result with the above market outcome.[9] The mandate m_j is understood also by the dealers. That is, dealers effectively observe m_j, and they still endogenously choose their symmetric service θ_j according to Proposition 1. Hence, for the rest of this section, we examine only the case of $\pi_j > \dot{\zeta}(0)$ to ensure that there is trading. Then the planner's objective function, the welfare expression w, remains as given in (14), subject to the symmetric θ_j, implied by (5). Denote by m_j^{P} the *p*lanner's optimal choice. To compare, denote by m_j^{M} the *m*arket outcome of customer j's dealer choice, as given in Corollary 1.

Proposition 5 (Planner's mandate of m_j). Assume (9) and (11). Then, welfare w is quasi-concave in m_j, and the planner's optimal choice m_j^{P} is unique in $[1, \hat{m}]$ and is always (weakly) lower than the market outcome: $m_j^{\mathrm{P}} \leq m_j^{\mathrm{M}}$. Specifically, let $h(\theta) := -(1-\theta)\ln(1-\theta)\dot{\zeta}(\theta) - \zeta(\theta)$. Then, if $\lim_{\theta\uparrow 1} h(\theta) < 0$, the planner always chooses $m_j^{\mathrm{P}} = 1$. If instead $\lim_{\theta\uparrow 1} h(\theta) \geq 0$, there exists a unique threshold $\theta^* \in (0, 1]$ such that $h(\theta^*) = 0$ and

(i) if $\dot{\zeta}(0) < \pi_j \leq \dot{\zeta}(\theta^*)$, then $m_j^{\mathrm{P}} = 1 < m_j^{\mathrm{M}}$;

(ii) if $\dot{\zeta}(\theta^*) < \pi_j < \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, then $1 < m_j^{\mathrm{P}} = 1 + \frac{\ln(\dot{\zeta}(\theta^*)/\pi_j)}{\ln(1-\theta^*)} < m_j^{\mathrm{M}} \leq \hat{m}$; and

(iii) if $\pi_j \geq \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, then $m_j^{\mathrm{P}} = m_j^{\mathrm{M}} = \hat{m}$.

We provide a heuristic discussion on why $m_j^{\mathrm{P}} \leq m_j^{\mathrm{M}}$. For simplicity, consider case (ii) above, where both m_j^{P} and m_j^{M} are interior, so that we can make use of the first-order derivatives to build intuition. The welfare expression (14) is the sum of the customer's trading gain and those of the m_j dealers: $w = \pi_j^{\mathrm{c}} + m_j\pi_j^{\mathrm{d}}$, where $\pi_j^{\mathrm{d}} = \theta_j \cdot \left(1-\theta_j\right)^{m_j-1}\pi_j - \zeta\left(\theta_j\right)$ follows (4). The planner's first-order

[9] We believe that mandating m_j is the most realistic and plausible policy intervention. In an electronic RFQ platform, mandating the m_j choice can be achieved by stipulating how many dealers a customer can reach in one "click." Although it does not happen in equilibrium, if a customer only chooses fewer dealers than stipulated, the platform could randomly select other dealers to fill the difference.

derivative is then

$$\frac{\mathrm{d}w}{\mathrm{d}m_j} = \frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} + \pi_j^{\mathrm{d}} + m_j \frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j}.$$

The first component, $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j}$, reflects the customer's consideration, as studied in Section 2.2.3. In particular, when choosing her optimal $m_j = m_j^{\mathrm{M}}$, unlike the planner, the customer does *not* internalize the following two effects on the dealers:

- The second term π_j^{d}, which is positive, reflects the marginal dealer's additional trading gain.
- The third term $m_j \frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j}$, which, rather intuitively, is always negative,[10] reflects the intensified competition among the dealers.

While the two effects are in opposite directions in general, we show in the proof that at the market outcome m_j^{M}, the negative competition effect dominates, i.e., $\pi_j^{\mathrm{d}} + m_j \frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j} < 0$. Intuitively, this is because at the customer's optimal m_j^{M}, the dealers' competition elasticity $\varepsilon(\theta_j)$ is necessarily very severe (Proposition 3), limiting their profit π_j^{d}. Not accounting for such dealer losses, the customer chooses a large optimal m_j^{M} to satisfy her first-order condition $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = 0$. Therefore, $\frac{\mathrm{d}w_j}{\mathrm{d}m_j}\Big|_{m_j = m_j^{\mathrm{M}}} < 0$, and the planner always wants to (locally) reduce her dealer choice m_j^{P} below the market outcome m_j^{M}.

Social planner mandating both m_j and $\{\theta_{ij}\}$. The social planner can also mandate dealers' service $\{\theta_{ij}\}$. Such regulations, though, might appear rather "invasive" as the planner has to interfere with how dealers run their businesses, and we do not consider such policies realistic. Nevertheless, for completeness, we briefly discuss this case below.

Note that from the planner's perspective, asking a customer not to contact certain dealers is the same as asking those dealers not to provide service to the customer. For example, if the planner wants customer j not to contact dealer i, forcing $i \notin \mathcal{D}_j$ is equivalent to requiring $\theta_{ij} = 0$. The planner's

[10] Indeed, $\frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j} = \frac{\partial \pi_j^{\mathrm{d}}}{\partial m_j} + \frac{\partial \pi_j^{\mathrm{d}}}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}$, but $\frac{\partial \pi_j^{\mathrm{d}}}{\partial \theta_j} = 0$ by the envelope theorem (as dealers choose their optimal θ_j). Hence, $\frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j} = \frac{\partial \pi_j^{\mathrm{d}}}{\partial m_j} = \theta_j \cdot (1-\theta_j)^{m_j - 1}\pi_j \ln(1-\theta_j) < 0$.

problem can then be rewritten as

$$\max_{\{\theta_{ij}\}}\left(1-\prod_{i=1}^{\hat{m}}(1-\theta_{ij})\right)\pi_j-\sum_{i=1}^{\hat{m}}\zeta(\theta_{ij}).$$

We say that a customer is *effectively* in business with $m_j^{\mathrm{P}}=\sum_{i=1}^{\hat{m}}\mathbb{1}_{\{\theta_{ij}>0\}}$ dealers.

Proposition 6 (Planner's mandate of both m_j and $\{\theta_{ij}\}$). Assume (9) and (11). Further, if $\varepsilon(\theta)>1$ for all $\theta\in[0,1]$, then the social planner chooses $m_j^{\mathrm{P}}=1$, so that each customer is effectively in business with at most one dealer.

Intuitively, when the competition elasticity is sufficiently high, choosing additional dealers results in all of them significantly reducing their service and lowering the trading probability. To avoid such inefficiency, the planner therefore chooses $m_j^{\mathrm{P}}=1$.

3.3 Market design recommendations

The above welfare analysis suggests that the market outcome is in general inefficient. In particular, since customers do not internalize dealers' competition cost, from a social planner's point of view, they reach out to "too many" dealers, whose lowered profitability can be socially costly. Allowing dealers to observe the customer's number of dealer contacts mitigates the excessiveness ($m_j^{\mathrm{M}}\le\hat{m}$) but does not fully address the issue ($m_j^{\mathrm{P}}\le m_j^{\mathrm{M}}$). Following Proposition 5, we now make two qualitative recommendations regarding the design of RFQ platforms.

First, the platform should make observable the number of dealers chosen by the customer.

Corollary 2 (Dealer competition observability). Following Proposition 5, welfare is weakly higher when dealers observe customers' m_j choice than when there is no such observability.

Proposition 4 has shown that customers are better off with such observability ($\pi_j^{\mathrm{c}}(m_j^{\mathrm{M}})\ge\pi_j^{\mathrm{c}}(\hat{m})$). So are the contacted dealers, because with the observability, the customers contact fewer dealers ($m_j^{\mathrm{M}}\le\hat{m}$), reducing their competition. The $\hat{m}-m_j^{\mathrm{M}}$ uncontacted dealers are worse off (because they no longer participate in trading), but they also no longer need to provide the costly service. Corollary 2 effectively shows that netting the above effects, welfare is always improved by the observability.

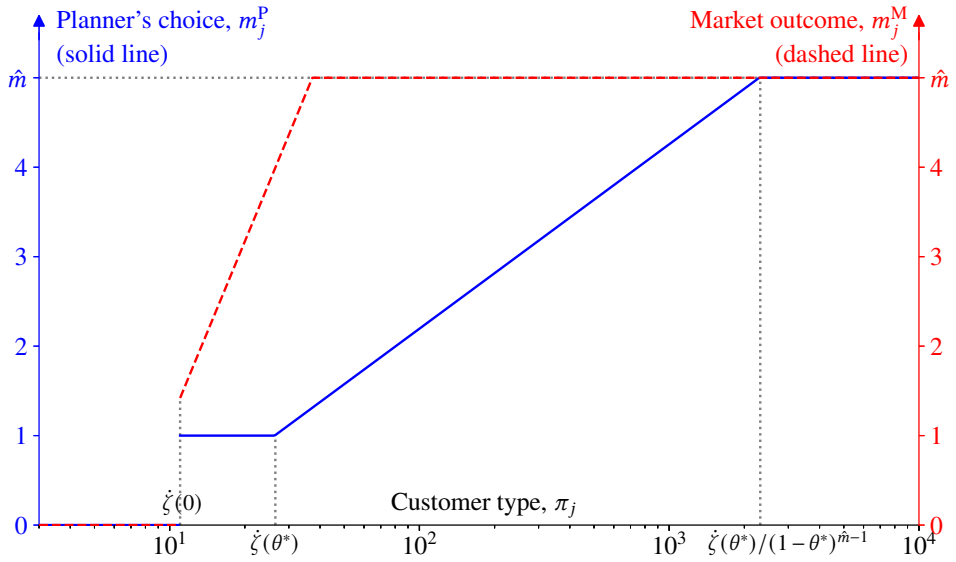


Figure 2: Social planner's welfare-maximizing dealer choice m_j^{P} vs. the market outcome m_j^{M}. This figure plots the planner's welfare-maximizing choice of m_j^{P} (solid line, left axis) and the customer's choice m_j^{M} (dashed line, right axis) as functions of the customer type π_j, varying in $\pi_j \in [3 \times 10^0, 10^4]$ on the horizontal axis (log scale). Dealers' service cost function $\zeta(\cdot)$ is parameterized as in (10), with $a = 1.0$, $b = 2.5$, and $c = 10.0$. The total number of dealers is set at $\hat{m} = 5$.

Second, the platform should restrict a customer's maximum number of dealer contacts, because the welfare-maximizing m_j^{P} is typically smaller than the market outcome m_j^{M}. Notably, different customers' trades should be subject to different restrictions:

Corollary 3 (Number of dealers and trade size). Following Proposition 5, the welfare-maximizing dealer choice m_j^{P} is weakly increasing in the trading gain size π_j.

Intuitively, since dealer service is (socially) costly, only large customers can justify the service costs from contacting more dealers. In practice, some RFQ platforms do impose a cap on the number of dealers a customer can contact: 5 on Bloomberg SEF (Riggs et al., 2020) and 4 on CanDeal (Allen and Wittwer, 2021). However, our model questions such a one-size-fits-all approach. Figure 2 illustrates the idea. The blue solid line indicates m_j^{P}, the socially optimal number of dealer contacts, while the red dashed line indicates the market outcome m_j^{M}. Following Proposition 5, m_j^{P} is always weakly lower than m_j^{M}, thus supporting the contact caps imposed by Bloomberg SEF and CanDeal. However, as the

trading gain size π_j increases, so does m_j^{P}, suggesting that if a customer enters a large trade size in the RFQ protocol, she should be allowed to contact more dealers. This is the market design idea implied by Corollary 3: A customer should be allowed to contact more dealers *only if* she wants to execute a sufficiently large position. If the trade size is small, the platform should limit her contact to contain the otherwise excessive dealer competition (and the socially wasteful dealer service cost).

It can be seen that both recommendations above aim to curb customers' excessive dealer contacting, which lowers dealers' expected profit. While in our model dealers' ex-ante participation $\hat{m}$ is exogenous, in more realistic settings, the lowered dealer profit can reflect in, for example, their reluctance in joining in RFQ platforms. This could contribute to the sluggish growth of electronic OTC trading, as evidenced by O'Hara and Zhou (2021). Our recommendations can alleviate the negative externality from customers to dealers, thus encouraging the latter's participation and improving efficiency.[11]

4 Endogenizing dealers' service cost

The previous analysis has assumed an exogenous dealer service cost $\zeta(\cdot)$. One natural source of such a cost is dealers' capacity constraints, like their computational power, limited labor force, and funding and inventory constraints, under which they will have to optimally allocate their limited capacity to serve different customers. This section studies such a model extension: Section 4.1 sets up the model, Section 4.2 characterizes the equilibrium, and Section 4.3 provides model predictions regarding dealer and customer behavior when the market is under stress.

[11] It should be noted, however, that the welfare improvement of our second recommendation is achieved at the cost of customers, whose endogenous participation in electronic platforms might be discouraged in a richer model environment. Transfers from dealers to customers, e.g., in the form of rebate to customers, can therefore offset such distributional inefficiency.

4.1 Model setup

We extend the setting of Section 2.1 by (i) introducing multiple customers and (ii) imposing a resource constraint on dealers' service. The details are discussed below.

Agents. We maintain the total number of homogeneous dealers as $\hat{m}$, the same as in Section 2.1. We then consider a continuum of customers of mass n (> 0), indexed by $j \in [0, n]$. Their types π_j, reflecting the total trading gains, can vary across j.

Finding dealers. Each customer j makes a dealer choice m_j as in Section 2.1.

Remark 6 (A continuum of customers). Since the dealers are homogeneous, each customer j randomly chooses to do business with m_j of them. Assuming a continuum of customers therefore helps ensure that every dealer receives almost surely the same amount of customers, so that dealers remain homogeneous. The customers can differ in their types π_j, reflecting different customers' urgency (willingness) to trade, the asset classes they specialize in, and/or their sophistication.

Dealers' service. Denote a dealer i's customers by $C_i \subset [0, n]$. As before, each dealer i observes both her customers' types π_j and their dealer choices m_j, for all $j \in C_i$. The dealer i then chooses her service $\theta_{ij} \in [0, 1]$ to every customer $j \in C_i$, subject to a resource constraint of

$$\int_{j \in C_i} \xi(\theta_{ij}) \mathrm{d}j \leq 1, \tag{15}$$

where $\xi(\cdot)$ translates the service θ_{ij} to the limited resource, and we normalize the endowment of this resource to be one unit. We assume that $\xi(\cdot)$ is convexly increasing, starting from $\zeta(0) = 0$, and thrice differentiable, with the first- and the second-order derivatives denoted, respectively, by $\dot{\xi}(\cdot)$ and $\ddot{\xi}(\cdot)$.

Remark 7 (Dealers' resource constraint). A dealer's resource constraint Equation (15) can arise for various reasons. First and foremost, time is limited. For example, it takes specifically trained traders to run time-consuming simulations to assess complicated structural products. If no pricing is obtained in time, the client might walk away for other options. Second, labor force is also limited. Experienced traders are few and maybe even fewer for the specific asset class that the client is interested. Risk

management staff are also important, as they approve or reject trades based on, for example, clients' creditworthiness, riskiness of trades, and the dealer's balance sheet. The back office is costly but necessary to run, owing to the heavy compliance and regulatory requirements. Third, the dealer's balance sheet capacity is limited. If inventory or capital has already been exhausted to facilitate other trades, a dealer will have to decline a client's request to trade.

Remark 8 (Cost functions $\zeta(\cdot)$ vs. $\xi(\cdot)$). Previously in Section 2, a dealer pays a cost of $\zeta(\theta_{ij})$, in dollars, to provide service θ_{ij} to customer j. In this section, there is no dollar cost in serving customers. Instead, each dealer is endowed with one unit of certain resource (e.g., time and/or labor), using which she can serve customers. The function $\xi(\cdot)$ translates the amount of service θ_{ij} into such limited resources. While $\xi(\theta_{ij})$ is not costly per se, as will be shown in Section 4.2.2, it implies a shadow cost to the dealer when the resource constraint binds. Such a shadow cost thus endogenizes the exogenous cost $\zeta(\cdot)$ assumed in Section 2.

Trading. The trading process remains as in Section 2.1.

Equilibrium. The three sets of equilibrium objects remain as in Section 2.1. In particular, we still focus on equilibria in which the homogeneous dealers use symmetric strategies, both in quoting to their customers and in choosing service θ_{ij} for a same customer j. The only difference is that now dealers need to account for the resource constraint (15) in optimizing their services $\{\theta_{ij}\}$.

4.2 Equilibrium analysis

As in Section 2.2, we analyze the equilibrium backwards. Much of the analysis remains the same as before, except that in studying dealers' service (Section 4.2.2), we will explicitly derive how dealers' resource constraint endogenizes the previously exogenous service cost $\zeta(\cdot)$.

4.2.1 Dealers' quoting

Given a symmetric service strategy, where every dealer $i \in \mathcal{D}_j$ provides the same service $\theta_{ij} = \theta_j$ to her customer j, the equilibrium quoting strategy in Section 2.2.1 remains the same. In particular, Lemmas 1 and 2 still hold.

4.2.2 Dealers' service to customers

Consider a dealer i. She observes $\{m_j, \pi_j\}$ for $j \in C_i$ and takes as given the competing dealers' symmetric service of $\theta_{i'j} = \theta_j$, $\forall i' \in \mathcal{D}_j$. Using (2), therefore, the dealer i's problem is

$$\max_{\theta_{ij}\in[0,1],\ \forall j\in C_i} \int_{j\in C_i} \theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1} \pi_j \mathrm{d}j, \quad \text{subject to } \int_{j\in C_i} \xi\left(\theta_{ij}\right)\mathrm{d}j \leq 1.$$

We assume for now that the capacity constraint will bind in equilibrium, i.e., $\int_{j\in C_i} \xi\left(\theta_{ij}\right)\mathrm{d}j = 1$, and later provide the necessary and sufficient condition in Section 4.2.4 for this assumption. The dealer's problem then has the following equivalent Lagrangian

$$\int_{j\in C_i} \theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1} \pi_j \mathrm{d}j - \kappa \cdot \left(\int_{j\in C_i} \xi\left(\theta_{ij}\right)\mathrm{d}j - 1\right), \tag{16}$$

where $\kappa\ (>0)$ is the shadow cost implied by the capacity constraint. Below we take κ as given and solve for dealers' symmetric service θ_j, until later in Section 4.2.4, where κ is pinned down in Lemma 5.

Endogenous cost $\zeta(\cdot)$. It can be seen from (16) that, *additively*, each customer $j \in C_i$ contributes

$$\left[\theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1} \pi_j - \kappa\xi\left(\theta_{ij}\right)\right]\mathrm{d}j \tag{17}$$

to dealer i's objective function. That is, in choosing the optimal service θ_{ij} to customer j, dealer i separately solves maximization problems for all $j \in C_i$, exactly as the problem (4) studied in Section 2.2.2. The only difference is that the previously exogenous service cost $\zeta(\cdot)$ now becomes

$$\zeta\left(\theta_{ij}\right) = \kappa\xi\left(\theta_{ij}\right), \tag{18}$$

with the *endogenous* resource shadow cost κ. Taking κ as given, dealers' symmetric service θ_j is still characterized by Proposition 1, with the cost function $\zeta(\cdot)$ given by (18).

4.2.3 Customers' choices of dealers

Taking the shadow cost κ (> 0) as given, then a customer j's optimization problem is exactly the same as in Section 2.2.3, with the cost function specified as in (18). Proposition 2 then holds, guaranteeing the existence of the optimal $m_j \in [0, \hat{m}]$. Note that using (18), the competition elasticity $\varepsilon(\cdot)$, as defined in (8), now becomes

$$\varepsilon(\theta_j) = \frac{1}{1-\theta_j} \frac{\kappa\dot{\xi}(\theta_j)}{\kappa\ddot{\xi}(\theta_j)} = \frac{1}{1-\theta_j} \frac{\dot{\xi}(\theta_j)}{\ddot{\xi}(\theta_j)}.$$

That is, following Section 2.3, as the key determinant of when the optimal $m_j \in (1, \hat{m})$, $\varepsilon(\theta_j)$ remains fully characterized by the exogenous function $\xi(\cdot)$, independent of κ. In particular, we shall continue to assume (9) and (11), so that Corollary 1 holds for those $\pi_j > \dot{\zeta}(0) = \kappa\dot{\xi}(0)$. (As before, if $\pi_j < \dot{\zeta}(0) = \kappa\dot{\xi}(0)$, this customer j never receives any service and is indifferent to choosing any m_j.)

4.2.4 Dealers' resource shadow cost

To summarize, thus far we have characterized the dealer's quoting strategies (in Section 4.2.1), their optimal symmetric service θ_j (in Section 4.2.2), and customers' optimal dealer choice m_j (in Section 4.2.3), *taking as given* dealers' resource shadow cost κ (> 0). To characterize the equilibrium, therefore, it remains to determine κ.

To do so, consider a dealer i. Since a customer $j \in [0, n]$ chooses to do business with m_j random dealers, the probability that i and j form a business pair is $\mathbb{P}[j \in C_i] = m_j/\hat{m}$. The dealer then provides service θ_j to customer j by spending $\xi(\theta_j)$ resources. Therefore, the dealer's resource constraint is

$$\int_{j \in C_i} \xi(\theta_{ij}) \mathrm{d}j = \int_0^n \frac{m_j}{\hat{m}} \xi(\theta_{ij}) \mathrm{d}j \leq 1. \tag{19}$$

The following lemma gives the exact parameter condition under which the above resource constraint

binds, so that $\kappa > 0$, as previously assumed in Section 4.2.2.

Lemma 5 (Shadow cost). Assume (11). Dealers' resource constraint (19) binds if and only if

$$n\xi(1) > 1, \tag{20}$$

under which the equality version of (19) uniquely determines the resource shadow cost κ (> 0).

To intuitively understand (20), note that if resource is unconstrained, then dealers will always provide maximum service $\theta_{ij} = 1$ for all customers, and every customer will choose the maximum number of $\hat{m}$ dealers. From the left-hand side of (19), the total resource spent in this case is $n\xi(1)$. Therefore, (20) simply ensures that the endowed unit of resource is insufficient for such maximum uses.

4.2.5 Summary of equilibrium

Summarily, assuming (9), (11), and (20) in this model extension, a dealer's service cost function $\zeta(\cdot)$ becomes $\kappa\xi(\cdot)$, where κ (> 0) is dealers' symmetric resource shadow cost and is uniquely determined by the binding resource constraint (19). The equilibrium is characterized by:

(i) every customer j contacts m_j dealers, where m_j is given in Corollary 1;

(ii) every dealer i provides symmetric service $\theta_{ij} = \theta_j$ as given in Proposition 1; and

(iii) every dealer i, if ready for customer j, quotes an ask price p_{ij} according to Lemma 1.

As discussed in Section 2.2.4, the equilibrium is unique up to all trading customers, i.e., those who have $\pi_j > \dot{\zeta}(0) = \kappa\dot{\xi}(0)$.

4.3 Predictions: Market in stress

To sharpen empirical predictions of the model, we examine, through the lens of our model, market stresses, such as downgrades of corporate bonds, the volatility in March 2020 due to COVID-19, and the market turmoil caused by UK's "mini-Budget," for example. To model such stress shocks, we consider the following parametrization of customer types $\{\pi_j\}$: A fraction $f_h \in [0, 1]$ of the mass-n customers are high-type with π_h, and the rest $f_l = 1 - f_h$ are low-type ($0 <$) with $\pi_l < \pi_h$. We

interpret the high-type as more urgent customers, hence with larger trading gain, than the low-type. The parametrization allows us to examine three different forms of market stress: larger n (more customers, lower per-capita dealer resource), higher f_h (larger fraction of urgent customers), and higher π_h (higher relative urgency). Although these shocks can all be thought of as market stress events, their implications can be rather different.

4.3.1 Larger n: More customers wanting to trade

One source of market stress is that increasingly more customers want to trade, especially in a short time frame, during which dealers' resource capacity cannot be easily adjusted and, hence, the resource available to each customer becomes smaller. We model such a shock via an exogenous increase in n, the total size of customers, and focus on the effects on the two sets of endogenous variables, the dealers' service allocation $\{\theta_h, \theta_l\}$, and the customers' dealer choices $\{m_h, m_l\}$. The results are summarized in the following proposition.

Proposition 7 (Market stress: Increased customer size, n). As the customer size n increases, both dealer service θ_j and customers' dealer choice m_j (weakly) decreases.

Figure 3(a)–(b) illustrate the patterns. It can be seen that dealers always provide less service to the low-type customers than to the high-type ($\theta_l < \theta_h$); and, knowing so, the low-type customers do business with fewer dealers than do the high-type ($m_l \leq m_h$). Further, as n increases, the lower per-capita resource limits dealers' service; hence, both θ_h and θ_l decrease with n. Notably, the low-type customers' service first drops to zero, at around $n \approx 20$, when the dealers find that their limited resource is too scarce to serve the less-profitable low-type customers. Consistently, from then on, $m_l = 0$—the low-type customers are "crowded out" for sufficiently large n.

Due to such a crowding-out effect, our model yields a novel empirical prediction that, during market stress times, the number of realized trades can be non-monotone in the severity of the stress. This result might be counterintuitive at first glance: Should customers not trade more aggressively when under stress, especially when there are more of them (larger n)? Our model highlights that,

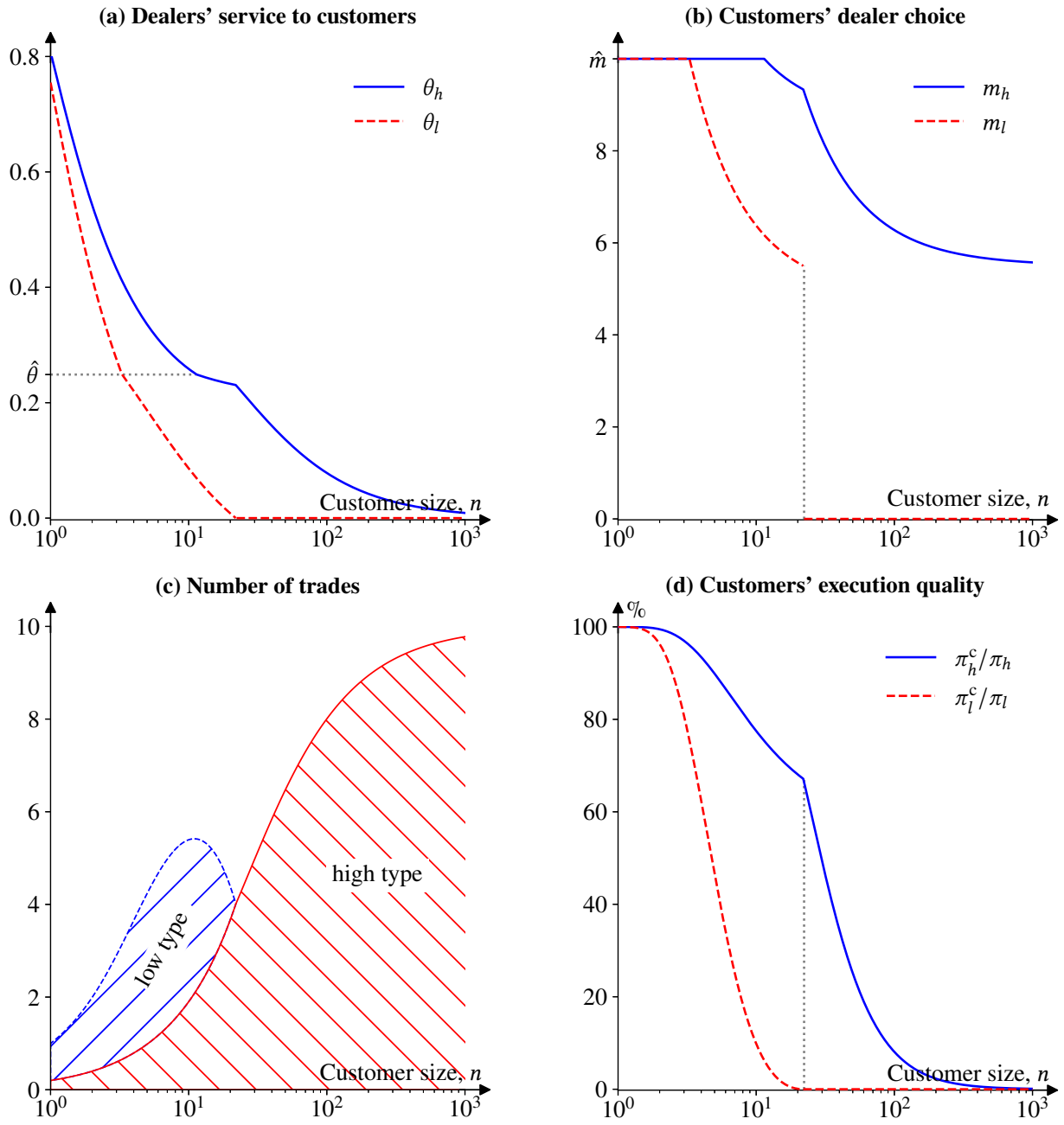


Figure 3: Market stress due to increased customer size, n. This figure plots how the customer size n, varying from $n = 1$ to $n = 10^3$, affects dealers' service in Panel (a), customers' dealer choice in (b), the number of trades in (c), and customers' execution quality in (d). There are two types of customers. A fraction of $f_h = 0.2$ of them have higher urgency to trade, with $\pi_h = 1$, and the rest $f_l = 0.8$ of them have $\pi_l = 0.1$. There is a total of $\hat{m} = 10$ dealers, and their service cost function $\xi(\cdot)$ is parameterized as in Example 1, with $a = 1.0$ and $b = 0.45$.

when dealers' service is constrained, not all customers will be served equally and some might be crowded out, creating nonmonotonicity.[12] The pattern is illustrated in Figure 3(c), where the number of low-type trades (the "//" patched area) initially increases with n but then quickly drops to zero (at around $n \approx 20$), thus creating a hump in the total number of trades. In contrast, the number of high-type trades (the "\\" patched area) increases with n, as the high-type is always served.

Figure 3(d) depicts customers' *execution quality*, measured as their expected trading gain as a percentage of the total trading gain, i.e., π_j^c/π_j for a type-j customer, where π_j^c follows (3) for $j \in \{l, h\}$. The measure is inspired by O'Hara, Wang, and Zhou (2018), who examine the execution quality of OTC trading by comparing the realized trading prices, and by Hendershott et al. (2022a), who demonstrate the importance of accounting for the probability of trading failure in measuring execution quality. Our measure nests both aspects, as reflected in (3). Consistent with the evidence from O'Hara, Wang, and Zhou (2018), our model predicts better execution quality for a more active customer (type-h, higher urgency), comparing the solid line with the dashed line. Further, as the stress exacerbates, the difference in the execution quality widens (until the low-types drop out).

4.3.2 Higher π_h: Relative urgency to trade

Market stress can alternatively take the form of an urgency shock on some customers. That is, some of the originally homogeneous customers might become more eager to trade, as reflected in their increased π_h ($> \pi_l$). Such a shock makes dealers more willing to spend their limited resource on serving the high-type customers, and, knowing this, the high-type customers also choose to do business with more dealers. Receiving the lower residual service, the low-type customers then contact fewer dealers.

[12] We recognize that the specific assumption of π_j matters for this effect. For example, if, instead, π_j is a smooth function of j, then the crowding out of the low-type customers will be smooth as well, and there will be no kink in Figure 3(c). However, the key underlying mechanism remains: certain low-type customers might be crowded out as dealers' resource constraint tightens.

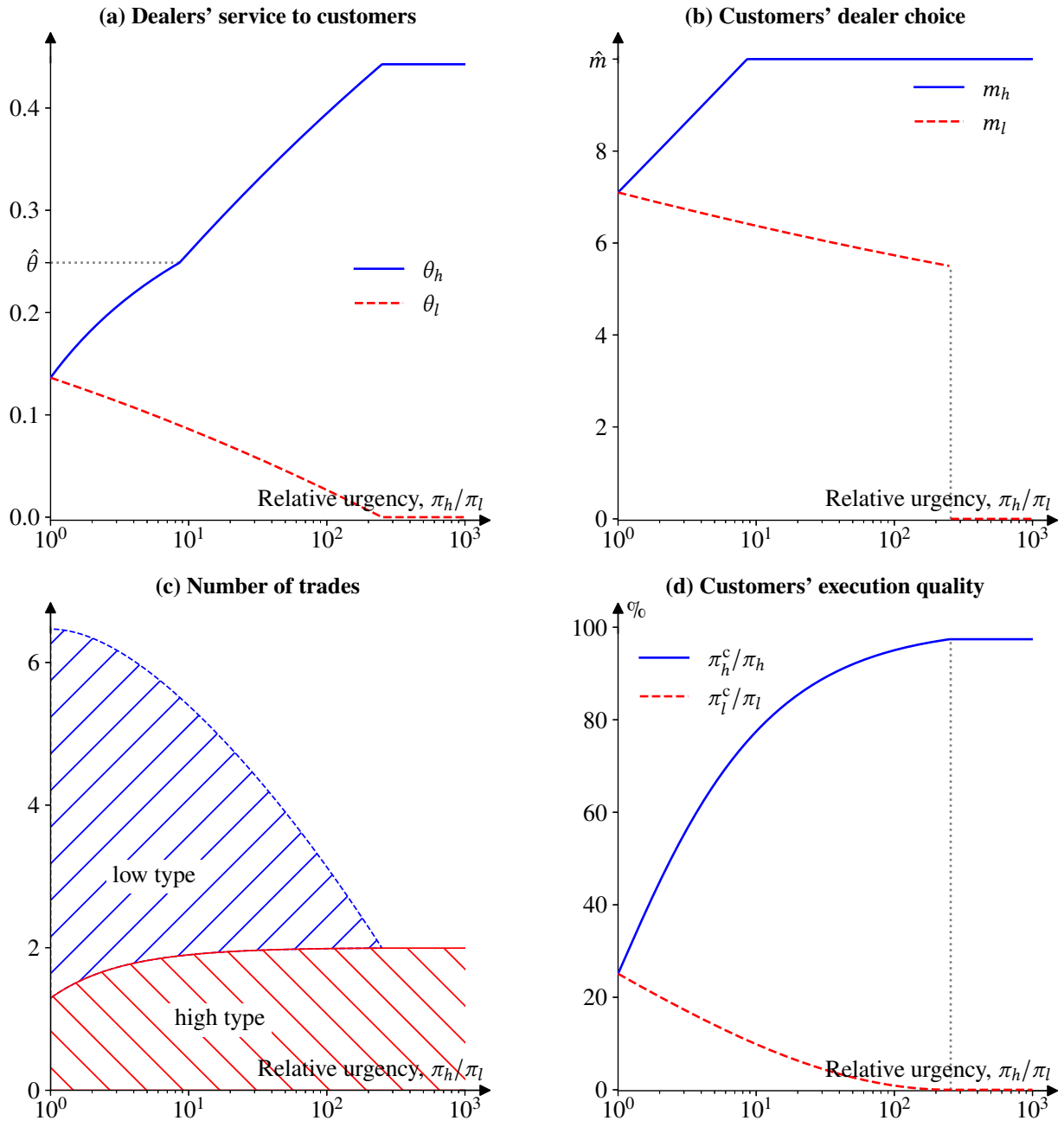


Figure 4: Market stress due to higher relative urgency, π_h/π_l. This figure plots how the urgency of high-type customers π_h, relative to the low-type π_l, varying from $\pi_h/\pi_l = 1$ to $\pi_h/\pi_l = 10^3$, affects dealers' service in Panel (a), customers' dealer choice in (b), number of trades in (c), and customers' execution quality in (d). There are two types of customers, with total mass $n = 10$. A fraction of $f_h = 0.2$ of them have higher urgency to trade, with π_h, and the rest $f_l = 0.8$ of them have $\pi_l = 1$. There is a total of $\hat{m} = 10$ dealers, and their service cost function $\xi(\cdot)$ is parameterized as in Example 1, with $a = 1.0$ and $b = 0.45$.

Proposition 8 (Market stress: Higher relative urgency, π_h/π_l). As π_h/π_l increases, the high-type (low-type) customers receive more (less) service and their dealer choices increase (decrease).

Figure 4(a)–(b) illustrate the patterns. Unlike the market stress seen in Figure 3 (increasing n), the relative urgency makes trading with the high-type customers more profitable, but less with the low-type less, for the dealers. Therefore, they cater to serving the high-types, who receive more service from and also reach out to more dealers (higher θ_h and m_h). In fact, if the relative profitability of the high-types becomes high enough ($\pi_h/\pi_l \approx 250$), the low-type customers completely drop out.

Figure 4(c) further illustrates that the crowding out of the low-type customers can be so severe that the overall trading can be hampered—less trading in more stressed times: The total number of trades (the sum of the "//" and the "\\" areas) decreases, at least initially, with the relative urgency π_h/π_l. Consistent with the above, Figure 4(d) shows that the high-type customers' execution quality continues to improve, at the cost of the low-types'.

4.3.3 More urgent customers, f_h

Yet another form of market stress is a shock that makes more customers feel urgent to trade, that is, an increase in the fraction f_h of high-type customers. Figure 5 illustrates the effects of such a shock. Notably, like the shock of an increase in π_h, the low-type customers are crowded out—they receive less service θ_l and also choose fewer dealers m_l—because dealers turn to serving the more profitable high-type customers. New to the shock in f_h, the high-type customers also receive less service and, hence, reach out to fewer dealers, i.e., both θ_h and m_h drop with f_h. This is because the high-type customers also compete against each other for dealers' limited resources. In other words, there is not only the inter-type crowding-out effect seen before, but also an *intra-type* crowding-out effect.

Proposition 9 (Market stress: A larger fraction of urgent customers, f_h). As f_h increases, both the high-type and the low-type customers receive less service and their dealer choices decrease.

Figure 5(c) illustrates the implication for trading activity. As more customers become high-type (more urgent to trade), the remaining low-type customers achieve fewer and fewer trades, not only

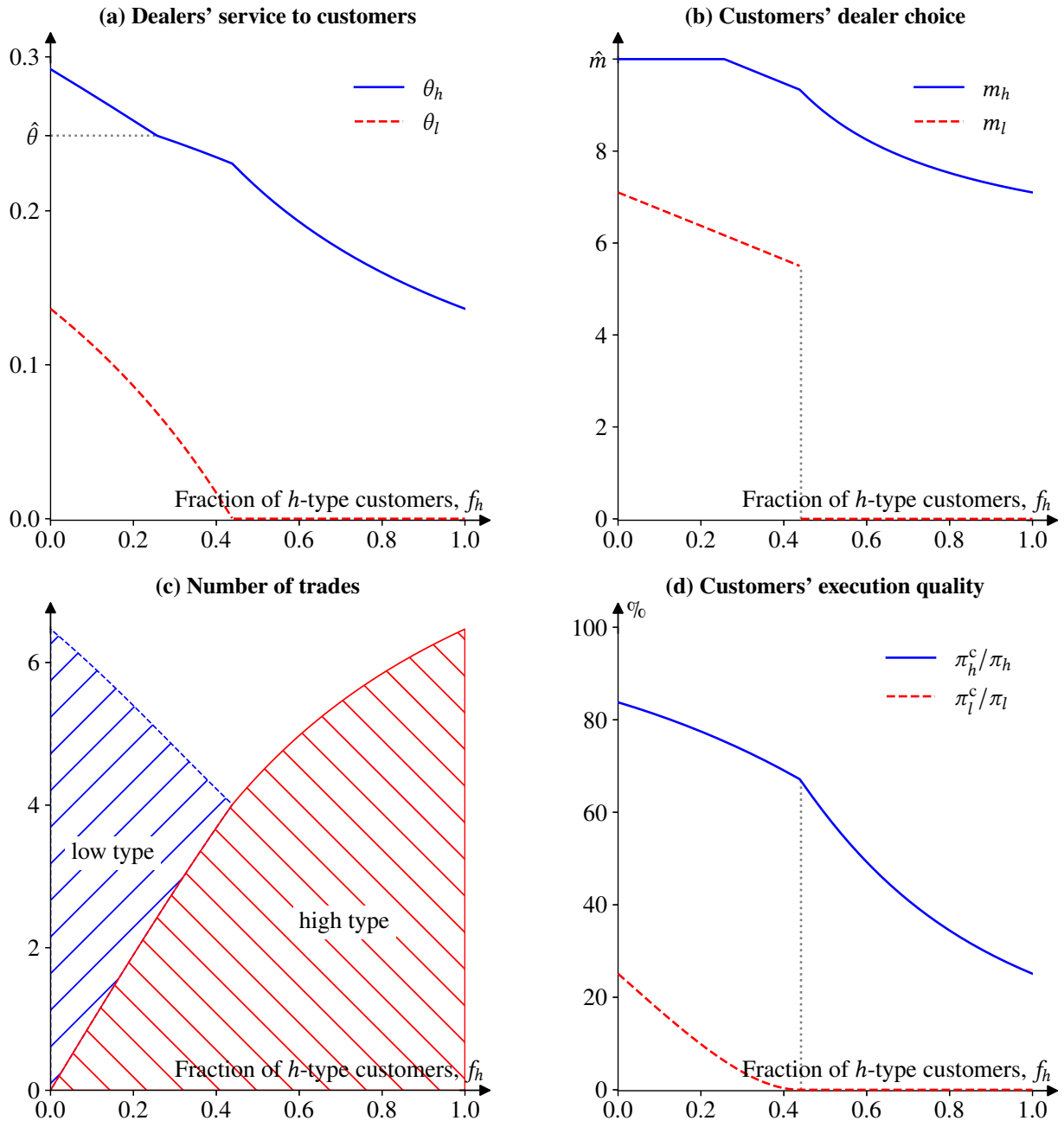


Figure 5: Market stress due to a larger fraction of urgent customers, f_h. This figure plots how the fraction of urgent customer f_h, varying from $f_h = 0$ to $f_h = 1$, affects dealers' service in Panel (a), customers' dealer choice in (b), number of trades in (c), and customers' execution quality in (d). There are two types of customers, with a total mass of $n = 10$. A fraction of f_h of them have higher urgency to trade, with $\pi_h = 1$, and the rest $f_l = 1 - f_h$ of them have $\pi_l = 0.1$. There is a total of $\hat{m} = 10$ dealers, and their service cost function $\xi(\cdot)$ is parameterized as in Example 1, with $a = 1.0$ and $b = 0.45$.

because $f_l = 1 - f_h$ decreases, but also because both θ_l and m_l are lower. On the other hand, the high-type customers in total trade more: Despite the fact that both θ_h and m_h decrease, making each of them less likely to trade, there are more of them as f_h increases. Together these two opposing effects generate the V-shaped pattern in aggregate.

Figure 5(d) shows that as the fraction of high-type customers increases, both types' execution quality deteriorates. This is again because of the crowding-out effect, both across types and within the h-type. Compared to the case of a relative urgent shock shown in Figure 4(d), it can be seen that depending on the nature of the market stress, the more urgent customers' execution quality can either improve or worsen with the severity of the stress.

5 Conclusion

This paper studies how customers choose their dealers in OTC trading. Muting the existing considerations (e.g., search costs, information concerns, and relationships), we develop a model and show that customers still refrain from exhausting all available dealers. The key friction lies in dealers' costly service to customers. Dealers then trade off such costs against the expected profit from trading, which is negatively affected by their competitors, i.e., the number of other dealers whom customers are contacting. Because of such a negative "service effect"—a novel mechanism emphasized in this paper—customers in equilibrium choose not to reach out to too many dealers. The model further speaks to regulation and market design issues in OTC trading. More over, model-implied empirical predictions speak to customer and dealer behavior during market stress periods.

Appendix

A Dealers' convex service cost

Section 2 assumes that the cost of dealer's service $\zeta(\cdot)$ is convex. This appendix shows that this assumption is without loss of generality: any $\zeta(\cdot)$ can be naturally "convexified" in our setting (and so

is the $\xi(\cdot)$ in Section 4).

Consider a dealer $i \in \mathcal{D}_j$, who needs to choose her service θ_{ij} to customer j. In doing so, she incurs a service cost of $\zeta(\theta_{ij}) : [0,1] \to \mathbb{R}^+$, which may or may not be convex. The dealer can play a mixed strategy with c.d.f. $G_{ij}(\theta_{ij})$ for $\theta_{ij} \in [0,1]$.

Suppose all other dealers in $\mathcal{D}_j$ play a symmetric strategy (possibly mixed) of $G_j(\cdot)$ with mean $\theta_j \in [0,1]$. It is easy to see that the analysis in Section 2.2.1 still goes through, and, in particular, both Lemma 1 and 2 hold: This is because a dealer i who is ready only cares about other dealers' probability of being ready, i.e., θ_j, the expectation of their possibly mixed strategy $\theta_{i'j}$ ($i' \in \mathcal{D}_j$ and $i' \neq i$).

Therefore, with the mixed strategy $G_{ij}(\cdot)$, dealer i's problem (4) becomes

$$\max_{G_{ij}(\cdot)} \bar{\theta}_{ij}\left(1-\theta_j\right)^{m_j-1}\pi_j - \int_0^1 \zeta(\theta_{ij})\mathrm{d}G_{ij}(\theta_{ij}),$$

where

$$\bar{\theta}_{ij} := \mathbb{E}[\theta_{ij}] = \int_0^1 \theta_{ij}\mathrm{d}G_{ij}(\theta_{ij})$$

is the dealer's expected amount of service under the mixed strategy $G_{ij}(\cdot)$. To solve the above problem, the dealer can proceed in the following two steps. First, she chooses a mixed strategy $G_{ij}(\cdot)$ to solve the following cost minimization problem, fixing any arbitrary expected service $\bar{\theta}_{ij} \in [0,1]$:

$$\bar{\zeta}(\bar{\theta}_{ij}) := \min_{G_{ij}(\cdot)} \int_0^1 \zeta(\theta_{ij})\mathrm{d}G_{ij}(\theta_{ij}), \text{ s.t. } \int_0^1 \theta_{ij}\mathrm{d}G_{ij}(\theta_{ij}) = \bar{\theta}_{ij}.$$

The minimized $\bar{\zeta}(\bar{\theta}_{ij})$ is the *effective cost function* of providing an expected amount of service $\bar{\theta}_{ij}$. Note that $\bar{\zeta}(\cdot)$ is by definition the lower boundary of the convex hull of the graph of $\zeta(\cdot)$ and therefore is a convex function in $\bar{\theta}$.[13] Note also that, while we began the analysis assuming the dealer is serving a specific customer j, the indirect cost $\bar{\zeta}(\cdot)$ does not depend on j.

Then in the second step, the dealer solves

$$\max_{\bar{\theta}_{ij}\in[0,1]} \bar{\theta}_{ij} \cdot \left(1-\theta_j\right)^{m_j-1}\pi_j - \bar{\zeta}(\bar{\theta}_{ij}),$$

which is identical to (4) studied in Section 2.2.2. Effectively, the above analysis shows that what matters is the "convexified" dealers' service cost $\bar{\zeta}(\cdot)$, and, hence, it is without loss of generality to assume that $\zeta(\cdot)$ is convex in the first place. Moreover, it follows immediately from the above analysis that when $\zeta(\cdot)$ is convex, it suffices to focus on pure strategies in θ_{ij}.

[13] The definition of $\bar{\zeta}(\cdot)$ is similar to the concept of concavification in **?** and is closely related to the notion of a biconjugate function in convex analysis (**?**).

B Proofs

Lemma 1

Proof. Consider first the trivial case of $m_j = 1$. There is then only one dealer in $\mathcal{D}_j$, who will always quote the highest possible price, i.e., the customer's reservation value π_j. This can be viewed as a degenerate mixed strategy with c.d.f. $F(\alpha)$ converging to a unity probability mass at $\alpha = 1$, as stated in the proposition.

Next consider $m_j \geq 2$. Without loss of generality, a dealer's strategy can be written as $\alpha\pi_j$ by choosing $\alpha \in [0, 1]$. Suppose α has a c.d.f. $F(\alpha)$ with possible realizations $[0, 1]$ (some of which might have zero probability mass). The following four steps pin down the specific form of $F(\cdot)$ so that it sustains a symmetric equilibrium.

Step 1: There are no probability masses in the support of $F(\cdot)$. If at $\alpha^* \in (0, 1]$ there is some non-zero probability mass, then any dealer has an incentive to deviate to quoting with the same probability mass but at a level infinitesimally smaller than α^*. In this way, she converts the strictly positive probability of tying with others at α^* to winning over them. (The undercut costs no expected revenue as it is infinitesimally small.) If at $\alpha^* = 0$ there is non-zero probability mass, again, any dealer who is ready will deviate, this time to an α just slightly above zero. This is because allocating probability mass at zero brings zero expected profit. Deviating to a slightly positive α, therefore, brings strictly positive expected profit. Taken together, there cannot be any probability mass in $\alpha \in [0, 1]$. Note that this rules out any pure symmetric-strategy equilibria.

Step 2: The support of $F(\cdot)$ is connected. The support is not connected if there is $(\alpha_1, \alpha_2) \subset [0, 1]$ on which there is zero probability assigned and there is probability density on α_1. If this is the case, then any dealer will deviate by moving the probability density on α_1 to any $\alpha \in (\alpha_1, \alpha_2)$. Such a deviation is strictly more profitable because doing so does not affect the probability of winning (if one wins at bidding α_1, she also wins at any $\alpha > \alpha_1$) and because $\alpha > \alpha_1$ is selling at a higher price.

Step 3: The upper bound of the support of $F(\cdot)$ is 1. The logic follows Step 2. Suppose the upper bound is $\alpha^* < 1$. Then, allocating the probability density at α^* to 1 is a profitable deviation: It does not affect the probability of winning and upon winning sells at a higher price.

Step 4: Deriving the c.d.f. $F(\cdot)$. Consider a specific dealer called i. Suppose all other dealers in $\mathcal{D}_j$, who are ready, quote according to some same distribution $F(\cdot)$. Quoting $\alpha\pi_j$, i gets to trade with the customer if, and only if, such a quote is the best. The customer examines all quotes received. For each of the $m_j - 1$ contacts, with probability $1 - \theta_j$ the dealer is not ready and in this case i's quote beats the no-quote. With probability θ_j, the contacted dealer is indeed ready and quotes at α'. Then, only with probability $\mathbb{P}(\alpha < \alpha') = 1 - F(\alpha)$ will i's quote win. Taken together, for each of the $m_j - 1$ potential

competitor, i wins with probability $\left(1-\theta_j\right)+\theta_j\cdot(1-F(\alpha))$, and he needs to win all these m_j-1 times to capture the trading gain of $\alpha\pi_j$. That is, i expects a profit of $\left(1-\theta_j F(\alpha)\right)^{m_j-1}\alpha\pi_j$. In particular, at the highest possible $\alpha=1$, the above expected profit simplifies to $\left(1-\theta_j\right)^{m_j-1}\Delta_{hd}$, because $F(1)=1$. In a mixed-strategy equilibrium, i must be indifferent to quoting any values of α in the support. Equating the two expressions above and solving for $F(\cdot)$, one obtains the c.d.f. stated in the proposition. It can then be easily solved that the lower bound of the support must be at $\left(1-\theta_j\right)^{m_j-1}$, where $F(\cdot)$ reaches zero. This completes the proof. □

Lemma 2

Proof. Given the mixed-strategy equilibrium, a dealer who is ready is indifferent to quoting any price $p_{ij}=\alpha_{ij}\pi_j$ when α_{ij} is in the support. In particular, by setting $\alpha_{ij}=1$, the expression in (2) is obtained. Since there are m_j such dealers, who each has a probability θ_j to be ready, they in total expect $m_j\theta_j\cdot(1-\theta_j)^{m_j-1}\pi_j$. The probability of trading is $1-(1-\theta_j)^{m_j-1}$. Therefore, the customer expects the residual (3). □

Proposition 1

Proof. The first-order derivative of (4) with respect to θ_{ij} is $\left(1-\theta_j\right)^{m_j-1}\pi_j-\dot{\zeta}(\theta_{ij})$, which, by symmetry of $\theta_j=\theta_{ij}$, becomes $\left(1-\theta_j\right)^{m_j-1}\pi_j-\dot{\zeta}(\theta_j)$ and is monotone decreasing in $\theta_j\in[0,1]$, owing to the assumed convexity of $\zeta(\cdot)$. Therefore, at the lower bound $\theta_j=0$, if the derivative is still negative, i.e., if $\pi_j\leq\dot{\zeta}(0)$, the optimal symmetric solution is $\theta_j=0$. At the upper bound $\theta_j=1$, the derivative evaluates to be $-\dot{\zeta}(1)<0$, implying that the optimal symmetric θ_j is never constrained from above. Hence, as long as $\pi_j>\dot{\zeta}(0)$, the first-order condition of $\left(1-\theta_j\right)^{m_j-1}\pi_j-\dot{\zeta}(\theta_j)$ implies a unique solution of $\theta_j=h(m_j,\pi_j)$. □

Proposition 2

Proof. Following Proposition 1, customers with $\pi_j\leq\dot{\zeta}(0)$ will only receive $\theta_j=0$. Hence, they are indifferent in their choices of m_j. Below we consider customers with $\pi_j>\dot{\zeta}(0)$, in which case dealers' first-order condition (5) holds and their optimal symmetric service $\theta_j=g(m_j,\pi_j)$, following Proposition 1. Note that the customer's objective π_j^{c}, as given in (3), is a function of both m_j and θ_j. Substituting $\theta_j=g(m_j,\pi_j)$, we then obtain a univariate optimization problem of $\max_{m_j\in[1,\hat{m}]}\pi_j^{\mathrm{c}}(m_j,\theta_j=g(m_j,\pi_j))$. Given the bounded support $[1,\hat{m}]$, an optimal m_j that maximizes π_j^{c} always exists. The optimal $m_j>1$ because at $m_j=1$, $\pi_j^{\mathrm{c}}=0$ (as, intuitively, the monopolist

dealer extracts all trading gain). By increasing to some $m_j > 1$, instead, the customer expects non-zero trading gain. □

Lemma 3

Proof. Directly evaluating the direct effect gives

$$\frac{\partial \pi_j^{\mathrm{c}}}{\partial m_j} = -(1-\theta_j)^{m_j-1}\big(\theta_j + (1-\theta_j+m_j\theta_j)\ln(1-\theta_j)\big)\pi_j. \tag{B.1}$$

Note that $m \geq 1$ and that $\ln(1-\theta) \leq 0$. Hence, the above is no smaller than $-(1-\theta_j)^{m_j-1}(\theta_j+\ln(1-\theta_j))$. Note further that $\theta_j \leq -\ln(1-\theta_j)$ for all $\theta_j \in [0,1)$. Therefore, the direct effect is weakly positive. Directly evaluating the indirect effect gives

$$\frac{\partial \pi_j^{\mathrm{c}}}{\partial \theta_j}\frac{\partial \theta_j}{\partial m_j} = (m_j-1)m_j \cdot (1-\theta_j)^{m_j-2}\theta_j \cdot \frac{\mathrm{d}\theta_j}{\mathrm{d}m_j},$$

which is weakly negative, because $m_j \geq 1$, $\theta_j \in [0,1]$, and $\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} \leq 0$ following (6). □

Proposition 3

Proof. We first show that if there is an interior solution of $m_j < \infty$, then $\varepsilon(\theta_j) > 2$. In this case, the customer's first-order condition $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = 0$ holds, i.e., following the analysis in the proof of Proposition 2,

$$\frac{(m_j-1)(\theta_j+(1-\theta_j)\ln(1-\theta_j))}{\theta_j+(1-\theta_j+m_j\theta_j)\ln(1-\theta_j)} + \frac{1}{\varepsilon(\theta_j)} = 0. \tag{B.2}$$

Define $v(x) := -x\ln(1-x)/(x+(1-x)\ln(1-x))$, which is increasing in $x \in (0,1)$ from $v(0) = 2$ to $\lim_{x\uparrow 1} v(x) = \infty$. Then rearrange (B.2) to get $\varepsilon(\theta_j) = (v(\theta_j)m_j-1)/(m_j-1) > (2m_j-1)/(m_j-1) > 2$, where the first inequality follows $v(\theta_j) > v(0) = 2$.

Consider now the sufficiency of $\varepsilon(0) > 2$. In the limit of $m_j \to \infty$, the θ_j implied by (5) converges to $\theta_j \to 0$. Then the left-hand side of (B.2) converges to $-\frac{1}{2} + 1/\varepsilon(\theta_j) < 0$. That is, in the limit of $m_j \to \infty$, π_j^{c} is decreasing. Therefore, there must exist some $m_j < \infty$ that maximizes π_j^{c}. □

Lemma 4

Proof. Following Proposition 1, customers with $\pi_j \leq \kappa\dot{\zeta}(0)$ will only receive $\theta_j = 0$. Hence, $\pi_j^c = 0$ for any m_j. Below we consider customers with $\pi_j > \kappa\dot{\zeta}(0)$. Evaluating the first-order derivative of (3) with respect to m_j yields that its sign is the same as the left-hand side of (B.2). Recall from dealers' first-order condition (5) and (6) that θ_j is a monotone decreasing function in m_j. Therefore, the left-hand

side of (B.2) can be seen as a function $f(\theta_j(m_j), m_j))$. Hence, at any stationary point $m_j^* \in (1, +\infty)$ (if exists), then $\text{sign}\left[\frac{\mathrm{d}^2\pi_C}{\mathrm{d}m_j^2}\right]\Big|_{m_j=m_j^*} = \text{sign}\left[\frac{\partial f}{\partial m_j} + \frac{\partial f}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}\right]\Big|_{m_j=m_j^*}$, where

$$\frac{\partial f}{\partial m_j} = \frac{(\theta_j + \ln(1-\theta_j))(\theta_j + (1-\theta_j)\ln(1-\theta_j))}{(\theta_j + (1-\theta_j+m_j\theta_j)\ln(1-\theta_j))^2} < 0,$$

for the numerator is negative (see the proof of Lemma 3); and

$$\frac{\partial f}{\partial \theta_j} = \frac{m_j(m_j-1)(\theta_j^2 - (1-\theta_j)(\ln(1-\theta_j))^2)}{(1-\theta_j)(\theta_j + (1-\theta_j+m_j\theta_j)\ln(1-\theta_j))^2} - \frac{1}{\varepsilon(\theta_j)^2}\frac{\mathrm{d}\varepsilon(\theta_j)}{\mathrm{d}\theta_j}.$$

It can be shown that $\theta_j^2 - (1-\theta)(\ln(1-\theta_j))^2$ is positive. Therefore, if $\frac{\mathrm{d}\varepsilon}{\mathrm{d}\theta_j} < 0$, then $\frac{\partial f}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} < 0$ and π_j^c is strictly concave at any stationary point. That is, π_j^c is a quasi-concave function in m_j. □

Corollary 1

Proof. **Existence and uniqueness.** The existence of the customer's optimal m_j follows Proposition 2. Under (11), π_j^c is quasi-concave in m_j, thus guaranteeing the uniqueness.

Monotonicity of m_j and θ_j in π_j. Accounting for the cap of $\hat{m}$, following the analysis in the proof of Proposition 3, the optimal m_j as a function of $\theta_j \in [0,1]$ can be written as

$$m_j = m(\theta_j) := \min\left\{\hat{m}, 1 + \frac{v(\theta_j) - 1}{\varepsilon(\theta_j) - v(\theta_j)}\right\}. \tag{B.3}$$

which is (weakly) increasing in θ_j. We now obtain two conditions, (5) and (B.3), for the two equilibrium objects $\{\theta_j, m_j\}$. Substituting (B.3) into (5) yields $(1-\theta_j)^{m(\theta_j)-1}\pi_j = \dot{\zeta}(\theta_j)$. The left-hand side is monotone decreasing, while the right-hand side is increasing in θ_j, thus yielding a unique solution of $\theta_j \in (0,1)$. Clearly, the implied θ_j is increasing in π_j. Therefore, the equilibrium $m_j = m(\theta_j)$ is also increasing in π_j. Recall from (B.3) that m_j is (weakly) increasing in θ_j. Hence, θ_j is also (weakly) increasing in π_j.

When m_j is interior. Following (B.3), m_j increases with θ_j but is capped by $\hat{m}$. By continuity, therefore, there is a unique threshold $\hat{\theta} \in (0,1)$ at which $m_j = \hat{m}$:

$$\hat{m} = 1 + \frac{v(\hat{\theta}) - 1}{\varepsilon(\hat{\theta}) - v(\hat{\theta})}. \tag{B.4}$$

That is, the optimal $m_j = \hat{m}$ if and only if the equilibrium $\theta_j \geq \hat{\theta}$, at which (5) becomes $(1-\hat{\theta})^{\hat{m}-1}\pi_j = \dot{\zeta}(\hat{\theta})$. Using the monotonicity above, therefore, $m_j < \hat{m}$ if and only if $\pi_j < \dot{\zeta}(\hat{\theta})/(1-\hat{\theta})^{\hat{m}-1}$. □

Proposition 4

Proof. Following Proposition 2, if $\pi_j \le \dot{\zeta}(0)$, then it does not matter whether the customer reveals m_j or not, as she never gets any service; i.e., $\pi_j^{\mathrm{c}}(m_j) = \pi_j^{\mathrm{c}}(\hat{m}) = 0$. Now suppose $\pi_j > \dot{\zeta}(0)$. Following Corollary 1, if the equilibrium $m_j = \hat{m}$, then $\pi_j^{\mathrm{c}}(m_j) = \pi_j^{\mathrm{c}}(\hat{m})$. If instead the endogenous optimal $m_j < \hat{m}$, then it follows that $\pi_j^{\mathrm{c}}(m_j) > \pi_j^{\mathrm{c}}(\hat{m})$. □

Proposition 5

Proof. **The shape of $w(m_j)$.** Welfare w as a function of m_j is given by (14). For now we ignore the constraint of $m_j \le \hat{m}$ and examine the whole support of $m_j \in [1, \infty)$ to characterize the shape of w. The first-order derivative is given by the $h(\cdot)$ function stated in the proposition:

$$\frac{\mathrm{d}w}{\mathrm{d}m_j} = \frac{\partial w}{\partial m_j} + \frac{\partial w}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} = -(1-\theta_j)^{m_j}\ln(1-\theta_j)\pi_j - \zeta(\theta_j) = h(\theta_j),$$

where the second equality holds because $\frac{\partial w}{\partial \theta_j} = m_j \cdot \Big((1-\theta_j)^{m_j-1}\pi_j - \dot{\zeta}(\theta_j)\Big) = 0$ following dealers' first-order condition (5); and the third equality makes use of (5) again by substituting $(1-\theta_j)^{m_j-1}\pi_j$. The second-order derivative then becomes $\frac{\mathrm{d}^2 w_j}{\mathrm{d}m_j^2} = \dot{h}(\theta_j)\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}$, where $\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} < 0$ following (6) and

$$\dot{h}(\theta_j) = \dot{\zeta}(\theta_j)\ln\big(1-\theta_j\big) - (1-\theta_j)\ddot{\zeta}(\theta_j)\ln\big(1-\theta_j\big) = -\ln\big(1-\theta_j\big)\dot{\zeta}(\theta_j)\left(\frac{1}{\varepsilon(\theta_j)} - 1\right).$$

It follows that $\frac{\mathrm{d}^2 w_j}{\mathrm{d}m_j^2} > 0$ if and only if $\varepsilon(\theta_j) > 1$. In particular, (5) implies that as m_j increases, θ_j eventually drops to $\lim_{m_j\to\infty}\theta_j = 0$, at which (9) ensures that $\varepsilon(0) > 2 > 1$. Also, $\lim_{m_j\to\infty}\frac{\mathrm{d}w}{\mathrm{d}\theta_j} = \lim_{\theta_j\to 0}\frac{\mathrm{d}w}{\mathrm{d}\theta_j} = 0$. Therefore, for sufficiently large m_j, w must be convexly decreasing. Then, following (11), for small m_j, w may be concave initially, before becoming convexly decreasing. In other words, w is quasi-concave in m_j. The quasi-concavity implies that the optimal m_j is uniquely determined by the first-order condition of $\frac{\mathrm{d}w}{\mathrm{d}m_j} = 0$, or $h(\theta_j) = 0$, if a *non-zero* solution of it exists.[14]

Suppose $h(\theta_j) = 0$ has a non-zero solution. Given the quasi-concavity, in this case, the non-zero solution uniquely maximizes w. Denote by $\theta^* \in (0, 1]$ the unique *non-zero* solution to $h(\theta_j) = 0$. Note that such a threshold θ^* is determined only by the shape of the service cost $\zeta(\cdot)$. Then following (5), the unconstrained optimal m_j is given by $m^* = 1 + \frac{\ln(\dot{\zeta}(\theta^*)/\pi_j)}{\ln(1-\theta^*)}$. However, the planner's optimal m_j^{P} is subject to the constraint of $m_j \in [1, \hat{m}]$. We then have two potential corners:

- If $m^* \le 1$, which is equivalent to $\pi_j < \dot{\zeta}(\theta^*)$, then $m_j^{\mathrm{P}} = 1$.
- If $m^* \ge \hat{m}$ (> 1), which is equivalent to $\pi_j > \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, then $m_j^{\mathrm{P}} = \hat{m}$.

[14] The first-order condition $h(\theta_j) = 0$ has a trivial solution of $\theta_j = 0$. But $\theta_j = 0$ produces the minimum welfare of zero (no dealer service) and, hence, cannot be optimal. We ignore this welfare-minimizing root to the first-order condition.

These two corners correspond to the cases (i) and (iii) in the proposition. Otherwise, i.e., when $1 < m^* < \hat{m}$, then $m_j^{\mathrm{P}} = m^*$ is interior, as stated in (ii).

Suppose $h(\theta_j) = 0$ has no non-zero solution. It is possible that $h(\theta_j) = 0$ does not have *non-zero* solution. In this case, the quasi-concavity, together with the fact that w decreases for sufficiently large m_j, implies that w is monotone decreasing in m_j, and, therefore, the optimal choice is $m_j^{\mathrm{P}} = 1$.

When does $h(\theta_j) = 0$ have no non-zero solution? Note that under (9) and (11), $h(\theta_j)$ initially decreases and may eventually increase in θ_j. Also, $h(0) = 0$. Therefore, there is no solution to $h(\theta) = 0$ if and only if $\lim_{\theta\uparrow 1} h(\theta) < 0$.

Comparison between $m_j^{\mathbf{P}}$ and $m_j^{\mathbf{M}}$. It remains to compare m_j^{P} with the market outcome m_j^{M}. To do so, we examine the marginal value of increasing m_j in the customer's problem and the planner's problem. Letting $\pi_j^{\mathrm{d}} = \frac{1}{m_j}(w_j - \pi_j^{\mathrm{c}})$ be the expected profit of each dealer, we have

$$\frac{\mathrm{d}(m_j\pi_j^{\mathrm{d}})}{\mathrm{d}m_j} = \frac{\mathrm{d}w_j}{\mathrm{d}m_j} - \frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = \dot{\zeta}(\theta_j)\left[\theta_j - \frac{\zeta(\theta_j)}{\dot{\zeta}(\theta_j)} + m_j\theta_j\ln(1-\theta_j)\frac{1/\varepsilon(\theta_j)}{m_j - 1 + 1/\varepsilon(\theta_j)}\right].$$

We evaluate $\frac{\mathrm{d}(m_j\pi_j^{\mathrm{d}})}{\mathrm{d}m_j}$ at the planner's unconstrained optimal choice of m_j^{P} (i.e., the m_j implied (5) at $\theta_j = \theta^*$). From the previous analysis, the corresponding θ^* satisfies $-(1-\theta^*)\dot{\zeta}(\theta^*)\ln(1-\theta^*) - \zeta(\theta^*) = 0$. Further, at θ^*, w must be locally concave and, hence, $\varepsilon(\theta^*) < 1$. Thus,

$$\begin{aligned}\frac{\mathrm{d}(m\pi_j^{\mathrm{d}})}{\mathrm{d}m_j}\Big|_{m_j=m_j^{\mathrm{P}}} =& \dot{\zeta}(\theta^*)\left[\theta^* + (1-\theta^*)\ln(1-\theta^*) + m_j^{\mathrm{P}}\theta^*\ln(1-\theta^*)\frac{1/\varepsilon(\theta^*)}{m_j^{\mathrm{P}} - 1 + 1/\varepsilon(\theta^*)}\right]\\ <& \dot{\zeta}(\theta^*)[\theta^* + (1-\theta^*)\ln(1-\theta^*) + \theta^*\ln(1-\theta^*)] = \dot{\zeta}(\theta^*)[\ln(1-\theta^*) + \theta^*] < 0.\end{aligned}$$

This shows that at the planner's unconstrained optimal mandate m_j^{P}, the customer has positive marginal value of increasing m_j. Therefore, the customer always chooses m_j^{M} weakly greater than m_j^{P}, with $m_j^{\mathrm{M}} > m_j^{\mathrm{P}}$ when $m_j^{\mathrm{P}} < \hat{m}$. □

Proposition 6

Proof. We prove the statement by contradiction. Suppose customer j is effectively in business with at least two dealers in the solution to the planner's problem. Let dealers 1 and 2 have $\theta_{1j} > 0$ and $\theta_{2j} > 0$. Note that it is never optimal for the planner to mandate any dealer to provide full service ($\theta_{ij} = 1$) and another dealer to provide positive service, since the planner can save cost without reducing expected trading gains by only keeping the dealer with full service. Therefore, both θ_{1j} and θ_{2j} are interior in

$(0,1)$, and they must satisfy the planner's first-order conditions, for $i \in \{1,2\}$:

$$\frac{\partial w}{\partial \theta_{ij}} = \prod_{k \neq i} \big(1-\theta_{kj}\big)\pi_j - \dot{\zeta}(\theta_{ij}) = 0.$$

Now we verify the second-order condition with respect to θ_{1j} and θ_{2j} by examining whether the Hessian matrix evaluated at θ_{1j} and θ_{2j} is negative (semi-)definite. We write down the sub-matrix and simplify it using the FOCs as follows,

$$\begin{aligned}
\begin{bmatrix} \frac{\partial^2 w_j}{\partial \theta_{1j}^2} & \frac{\partial^2 w_j}{\partial \theta_{1j}\partial \theta_{2j}} \\ \frac{\partial^2 w_j}{\partial \theta_{1j}\partial \theta_{2j}} & \frac{\partial^2 w_j}{\partial \theta_{2j}^2} \end{bmatrix} &= \begin{bmatrix} -\ddot{\zeta}(\theta_{1j}) & -\Pi_{k>2}(1-\theta_{kj})\pi_j \\ -\Pi_{k>2}(1-\theta_{kj})\pi_j & -\ddot{\zeta}(\theta_{2j}) \end{bmatrix}, \\
&= \begin{bmatrix} -\ddot{\zeta}(\theta_{1j}) & -\dot{\zeta}(\theta_{1j})/(1-\theta_{2j}) \\ -\dot{\zeta}(\theta_{2j})/(1-\theta_{1j}) & -\ddot{\zeta}(\theta_{2j}) \end{bmatrix}.
\end{aligned}$$

Next we calculate the determinant of the matrix,

$$\begin{vmatrix} \frac{\partial^2 w_j}{\partial \theta_{1j}^2} & \frac{\partial^2 w_j}{\partial \theta_{1j}\partial \theta_{2j}} \\ \frac{\partial^2 w_j}{\partial \theta_{1j}\partial \theta_{2j}} & \frac{\partial^2 w_j}{\partial \theta_{2j}^2} \end{vmatrix} = \ddot{\zeta}(\theta_{1j})\ddot{\zeta}(\theta_{2j}) - \frac{\dot{\zeta}(\theta_{1j})\dot{\zeta}(\theta_{2j})}{(1-\theta_{1j})(1-\theta_{2j})} = \ddot{\zeta}(\theta_{1j})\ddot{\zeta}(\theta_{2j})\big[1-\varepsilon(\theta_{1j})\varepsilon(\theta_{2j})\big] < 0.$$

The last inequality holds because $\ddot{\zeta}(\cdot) > 0$ and $\varepsilon(\cdot) > 1$ for any $\theta \in (0,1)$. The negative determinant indicates that the matrix is not negative semi-definite. Thus, θ_{1j} and θ_{2j} do not satisfy the second-order condition, and therefore cannot form a local maximum. The contradiction shows that there is at most one dealer providing service to the customer if the planner mandates both $\{\theta_{ij}\}$ and m_j. □

Corollary 2

Proof. Given (9) and (11), Proposition 5 shows that: a) welfare w_j is quasi-concave in m_j and b) $m_j^P \leq m_j^M \leq \hat{m}$. It follows immediately that welfare is weakly decreasing in m_j between m_j^P and $\hat{m}$. Therefore, the welfare at $m_j = m_j^M$ (when m is observable) is weakly higher than the welfare at $m_j = \hat{m}$ (when m is unobservable). □

Corollary 3

Proof. This is a direct implication of Proposition 5. When $\dot{\zeta}(0) < \pi_j \leq \dot{\zeta}(\theta^*)$, $m_j^{\mathrm{P}} = 1$. When $\dot{\zeta}(\theta^*) < \pi_j < \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, $m_j^{\mathrm{P}} = 1 + \frac{\ln(\dot{\zeta}(\theta^*)/\pi_j)}{\ln(1-\theta^*)}$ increases from 1 to $\hat{m}$. When $\pi_j \geq \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, $m_j^{\mathrm{P}} = \hat{m}$. □

Lemma 5

Proof. **Monotonicity of m_j and θ_j in κ.** Given κ, dealer i chooses $\{\theta_{ij}\}$ to maximize

$$\theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1}\pi_j - \kappa\xi\left(\theta_{ij}\right),$$

and customer j chooses m_j to maximize

$$\pi_j^{\mathrm{c}} := \left(1-(1-\theta_j)^{m_j} - m_j\theta_j \cdot (1-\theta_j)^{m_j-1}\right)\pi_j.$$

Note that if we replace π_j with π_j/κ and κ with 1, the optimal m_j and θ_j remains the same. Therefore, The effect of an increase in κ on m_j and θ_j is isomorphic to a decrease in π_j. In the proof of Corollary 1, we have shown that m_j and θ_j continuously increases in π_j. This implies that that both m_j and θ_j ($j = h, l$) continuously decrease in κ.

Uniqueness of κ. When (19) binds, it implies at most one solution of κ. This is because, given that both m_j and θ_j are monotone decreasing in κ, so is the left-hand side of (19).

Existence of κ. Next, we characterize when a solution of $\kappa > 0$ exists. On the one hand, in the upper limit of $\kappa \uparrow \infty$, there is clearly no service from any dealer i for any customer j, i.e., $\theta_{ij} = 0$: dealers' first-order condition (16) fails for any $\theta_j > 0$. Then $\zeta(\theta_j) \to \zeta(0) = 0$ and $\int_0^n \frac{m_j}{\hat{m}}\xi(\theta_{ij})\mathrm{d}j \to 0 < 1$, for any $n > 0$ (because $m_j < \hat{m} < \infty$). On the other hand, if $\kappa \downarrow 0$, then (5) implies $\theta_j \uparrow 1 > \hat{\theta}$, $m_j \uparrow \hat{m}$ (Proposition 2), and hence, $\int_0^n \frac{m_j}{\hat{m}}\xi(\theta_{ij})\mathrm{d}j \to n\zeta(1)$. Therefore, there is a unique solution of $\kappa > 0$ if and only if $n\zeta(1) > 1$. □

Proposition 7

Proof. Under the parametrization in Section 4.3, the dealers' resource constraint (19) becomes

$$\int_{j\in C_i} \xi(\theta_{ij})\mathrm{d}j = \frac{n}{\hat{m}}\left[f_h m_h \xi(\theta_h) + n f_l m_l \xi(\theta_l)\right] = 1. \tag{B.5}$$

In the proof of Lemma 5, we have shown that both m_j and θ_j are decreasing in κ. Thus, the left-hand side of (B.5) is decreasing in κ. To sustain the resource constraint (B.5), an increase in n must correspond to an increase in the dealers' shadow cost κ, and thus both m_j and θ_j decrease, $j \in \{l, h\}$. □

Proposition 8

Proof. In the proof of Lemma 5 we have shown that both m_j and θ_j increase in π_j/κ. Therefore, if we focus on the two-type parametrization, the binding resource constraint (B.5) implies that π_h/κ and π_h/κ must move in different directions when π_h/π_l increases. Note that $\pi_h/\pi_l = (\pi_h/\kappa)/(\pi_l/\kappa)$. It

follows immediately that an increase in π_h/π_l leads to an increase in π_h/κ and a decrease in π_h/κ, and thus an increase (decrease) in m_h and θ_h (m_l and θ_l). □

Proposition 9

Proof. We have already shown that the left-hand side of (B.5) is decreasing in κ (Proposition 7). Also note that the left-hand side of (B.5) is increasing in f_h. To keep the resource constraint (B.5) hold, an increase in f_h must correspond to an increase in the dealers' shadow cost κ, and thus a decrease in both m_j and θ_j ($j = h, l$). □

References

Allen, Jason and Milena Wittwer. 2021. "Centralizing over-the-counter markets?" Working paper.

Baldauf, Markus and Joshua Mollner. 2022. "Competition and Information Leakage." Working paper.

Bernhardt, Dan, Vladimir Dvoracek, Eric Hughson, and Ingrid M. Werner. 2005. "Why Do Larger Orders Receive Discounts on the London Stock Exchange?" *The Review of Financial Studies* 18 (4):1343–1368.

Bessembinder, Hendrik, Chester Spatt, and Kumar Venkataraman. 2020. "A Survey of the Microstructure of Fixed-Income Markets." *Journal of Financial and Quantitative Analysis* 55 (1):1–45.

Breckenfelder, Johannes, Pierre Collin-Dufresne, and Stefano Corradin. 2022. "Is the bond market competitive? Evidence from the ECB's asset purchase programme." Working paper.

Burdett, Kenneth and Maureen O'Hara. 1987. "Building Blocks: An Introduction to Block Trading." *Journal of Banking and Finance* 11 (2):193–212.

Collin-Dufresne, Pierre, Peter Hoffmann, and Sebastian Vogel. 2022. "Informed Traders and Dealers in the FX Forward Market." Working paper.

Desgranges, Gabriel and Theirry Foucault. 2005. "Reputation-based pricing and price improvements." *Journal of Economics and Business* 57 (6):493–527.

Duffie, Darrell, Piotr Dworczak, and Haoxiang Zhu. 2017. "Benchmarks in Search Markets." *The Journal of Finance* 72 (5):1983–2044.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2005. "Over-the-Counter Markets." *Econometrica* 73 (6):1815–1847.

Glebkin, Sergei, Bart Zhou Yueshen, and Ji Shen. 2022. "Simultaneous Multilateral Search." *The Review of Financial Studies* Forthcoming.

Glode, Vincent and Christian C. Opp. 2020. "Over-the-Counter versus Limit-Order Markets: The Role of Traders' Expertise." *The Review of Financial Studies* 33 (2):866–915.

Hammermann, Felix, Kieran Leonard, Stefano Nardelli, and Julian von Landesberger. 2019. "Taking stock of the Eurosystems asset purchase programme after the end of net asset purchases." Technical report.

Hau, Harald, Peter Hoffmann, Sam Langfield, and Yannick Timmer. 2021. "Discriminatory Pricing of Over-the-Counter Derivatives." *Management Science* 67 (11):6660–6677.

Hendershott, Terrence, Dan Li, Dmitry Livdan, and Norman Schürhoff. 2020. "Relationship Trading in OTC Markets." *The Journal of Finance* 75 (2):683–734.

———. 2022a. "True Cost of Immediacy." Working paper.

Hendershott, Terrence, Dan Li, Dmitry Livdan, Norman Schürhoff, and Kumar Venkataraman. 2022b. "Quote Competition in Corporate Bonds." Working paper.

Hendershott, Terrence and Ananth Madhavan. 2015. "Click or Call? Auction versus Search in the Over-the-Counter Market." *The Journal of Finance* 70 (1):419–447.

Hollifield, Burton, Artem Neklyudov, and Chester Spatt. 2017. "Bid-Ask Spreads, Trading Networks, and the Pricing of Securitizations." *The Review of Financial Studies* 30 (10):3048–3085.

Jovanovic, Boyan and Albert J. Menkveld. 2022. "Equilibrium Bid-Price Dispersion." *Journal of Political Economy* 130 (2):426–461.

Kondor, Péter and Gábor Pintér. 2022. "Clients' Connections: Measuring the Role of Private Information in Decentralized Markets." *The Journal of Finance* 77 (1):505–544.

Levin, Dan and James L Smith. 1994. "Equilibrium in auctions with entry." *The American Economic Review* :585–599.

Li, Dan and Norman Schurhoff. 2019. "Dealer Networks." *The Journal of Finance* 74 (1):91–144.

Li, Wei and Zhaogang Song. 2021. "Dealer Expertise and Market Concentration in OTC Trading." Working paper.

Liu, Ying, Sebastian Vogel, and Yuan Zhang. 2017. "Electronic Trading in OTC Markets vs. Centralized Exchange." Working paper.

Maggio, Marco Di, Amir Kermani, and Zhaogang Song. 2017. "The value of trading relations in turbulent times." *Journal of Financial Economics* 124 (2):266–284.

Menezes, Flavio M and Paulo K Monteiro. 2000. "Auctions with endogenous participation." *Review of Economic Design* 5 (1):71–89.

O'Hara, Maureen, Yihui Wang, and Xing (Alex) Zhou. 2018. "The execution quailty of corporate bonds." *Journal of Financial Economics* 130 (2):308–326.

O'Hara, Maureen and Xing Zhou. 2021. "The Electronic Evolution of Corporate Bond Dealers." *Journal of Financial Economics* 140 (2):368–390.

Pinter, Gabor, Chaojun Wang, and Junyuan Zou. 2022. "Information chasing versus adverse selection." Working paper.

Riggs, Lynn, Esen Onur, David Reiffen, and Haoxiang Zhu. 2020. "Swap Trading after Dodd-Frank:

Evidence from Index CDS." *Journal of Financial Economics* 137 (3):857–886.

Stigler, George J. 1961. "The Economics of Information." *Journal of Political Economy* 61 (3):213–225.

Vogel, Sebastian. 2019. "When to Introduce Electronic Trading Platforms in Over-the-Counter Markets?" Working paper.

Wang, Chaojun. 2022. "The Limits of Multi-Dealer Platforms." Working paper.