Essays in Market Microstructure

by

Ariel Lohr

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2022 by the
Graduate Supervisory Committee:

Hendrik Bessembinder, Chair
Sunil Wahal
George Aragon

ARIZONA STATE UNIVERSITY

May 2022

ABSTRACT

This dissertation consists of two essays. The first, titled "Sweep Order and the Cost of Market Fragmentation" takes a "revealed-preference" approach towards gauging the effects of market fragmentation by documenting the implicit costs borne by traders looking to avoid executing in a fragmented environment. I show that traders use Intermarket Sweep Orders (ISO) to trade "as-if" markets were single-venued and pay a premium to do so. Using a sample of over 2,600 securities over the period January 2019 to April 2021, this premium amounts to 1.3 bps on average (or 40% of the effective spread), amounting to a total of $3 billion over the sample period. I find a positive, robust, and significant relationship between the premium and different measures of market fragmentation, further supporting the interpretation of the premium as a cost of market fragmentation. The second essay, titled "The Profitability of Liquidity Provision" investigates the relationship between the profits realized from providing liquidity and the amount of time it takes liquidity providers to reverse their positions. By tracking the cumulative inventory position of all passive liquidity providers in the US equity market and matching each aggregate position with its offsetting trade, I construct a measure of profits to liquidity provision (realized profitability) and assess how profitability varies with the average time to offset. Using a sample of all common stocks from 2017 to 2020, I show that there is substantial variation in the horizon at which trades are turned around even for the same stock. As a mark-to-market profit, the conventional realized spread—measured with a prespecified horizon—can deviate significantly from the realized profits to liquidity provision both in the cross-section and in the time-series. I further show that, consistent with the risk-return tradeoff faced by liquidity providers as a whole, realized profitability is low for trades that are quickly turned around and high for trades that take longer to reverse.

ACKNOWLEDGEMENT

I wish to express my great debt of gratitude to my doctoral committee of Hendrik Bessembinder, Sunil Wahal, and George Aragon. At each step of the process they provided me with the support and guidance without which my work would doubtlessly be lacking.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

SWEEP ORDERS AND THE COST OF MARKET FRAGMENTATION

1.1 Introduction

Over the last 20 years the public security markets in the United States have fragmented. At the turn of the century individual securities were almost exclusively traded on a single exchange. Today that same secutiry may be traded across as many as 13 different public exchanges[1]. For instance Menkveld (2013) reports that the NYSE's market share in NYSE-listed stocks fell from 80% in 2005 to 25% in 2010. Whether such fragmentation is a good or bad development is the subject of open discussion in the theoretical literature. A fragmented market is said to lower transaction costs through increased cross-exchange competition on factors such as fees, rules, and access to data. On the other hand, a single-venued market, benefiting from the reduced complexity of a common order book, allows for a quicker and more certain trade-fill. To this end the SEC implemented Regulation National Market System (Reg NMS) in 2007 with the expressed intent of capturing the benefits of an integrated market while preserving the gains from cross-exchange competition. Recent discussion regarding a "Reg NMS II"[2] demonstrates a continued interest on the part of both regulators and practitioners in the costs of a fragmented market. I contribute to this discussion by estimating the costs that traders are willing to bear in order to avoid the delays and uncertainty of a fragmented environment and trade as-if there were a single consolidated order-book.

[1] Not to mention the dozens of off-exchange alternative trading venues also available

[2] https://www.sec.gov/rules/proposed/2020/34-88216.pdf

https://www.nasdaq.com/docs/2020/05/27/Reg-NMS-II-Comment-Letter.pdf

In modern markets, a prohibition on trade-throughs has meant that a market order must execute at a price no worse than the best protected quote[3] (i.e. the NBBO) available on any market.[4] It's execution may therefore be delayed as its routing is determined, and particularly so if the size of the order is greater than the quantity initially available at the NBBO quote. In this case the NBBO must be updated before the remainder of the order can be executed. Further, order cancellations in the wake of the initial partial execution can cause the new NBBO quote to become less advantageous to the market order. In contrast, an intermarket sweep order (ISO) is immediately executed on one or more exchanges, requiring only that the order submitter simultaneously "sweep" NBBO orders resting on other exchanges at the time of order submission. As a consequence, the ISO can execute at an average price worse than the NBBO. An ISO implemented trade is free to "walk to the book" on one exchange if it also clears out any better-priced protected quotes, meaning that it only interacts with a subset of the total available liquidity. As a result, it is possible that some better priced unprotected liquidity on another exchange could be missed by the ISO. What an ISO offers in return to missing liquidity is the ability to instantaneously trade against that subset of quotes as would be the case on a consolidated book with similar liquidity; in other words, ISOs trade as-if markets were not fragmented. Figure 1 demonstrates this dynamic.

In this study I take a "revealed-preference" approach towards measuring the costs of market fragmentation by comparing the executions of ISO and non-ISO trades. The ideal comparison is to compare the costs of each ISO trade to what they would've been had a sequence of MOs been used instead, but this is not possible as the counterfactual

[3] An exchange's "protected" quotes correspond to the highest priced Bid and lowest priced Ask quotes visible on that exchange.

[4] A trade-through refers to when a marketable order is filled at a price inferior to the prevailing NBBO,Section §242.611 of the SEC's Regulation National Market System

trades are not observable. However, because MOs should execute at the NBBO, which is observable, it is possible to calculate what each trade's fill would've been had they interacted with the NBBO first. As an example, suppose that, out of a sequence of ISOs, we observe a sale order for 100 shares being cleared at a price of \$3.5 even though the updated NBB quotes 50 shares at \$4.5. If the trade first interacted with the NBB, 50 shares would have sold at \$4.5 before the next 50 shares being off-loaded at \$3.5. In total, the MO trading could have earned the trader an extra \$50 (\$400 vs \$350) or $\frac{\$50}{100} = \$0.50$ per share (if the quotes did not change). The presumption here is that, the trader, by opting for the ISO was willing to pay at least \$0.50 per share to avoid trading with MOs.

I consider a sample of over 2,800 securities traded on US public exchanges for the time period spanning January 2019 to April 2021. Within this sample, which makes up 93% of traded dollar volume, ISOs accounted for 48% of on-exchange trade-volume. ISOs on average executed at prices 1.36 basis points (about $\frac{1}{5}$ of the quoted spread) worse than similar non-ISO trades. My central argument is that, since traders make use of ISOs to trade as-if markets were not fragmented, the difference in costs reflects the premium paid by traders to quickly trade in an integrated environment. To put a dollar value on this premium, over \$90 million in ISO premiums were paid by traders in the SPY market alone. When considering the whole sample of securities, the premiums amounted to \$2.97 billion for this same time period.

I find that the ISO premium is higher in more fragmented markets and lower in less fragmented ones, this is inline with the interpretation of the premium as the price of avoiding a fragmented market. In absolute terms, the ISO premium is higher for stocks traded across 4 or less venues (5.22 bps) than for securities traded on more than 9 venues (1.51 bps). This difference is largely driven by the fact that transaction costs, as a whole, are large for small stocks than for large, and large

stocks are more likely to be traded across more exchanges. Measured as a fraction of effective spreads, the premium of 0.42 spreads for securities traded on many exchanges is greater than the 0.32 for securities traded on few exchanges. This is not to say that the variation in the premium is caused by the number of trading venues, something which is highly endogenous Bessembinder (2003). Rather, I am documenting that the (scaled) premium is larger in environments where one would expect fragmentation to be a more pronounced issue.

Regression analysis provides further evidence of a statistically and economically significant positive relationship existing between the size of the premium and the extent to which markets are fragmented. I find that a 1% increase in trade fragmentation, as measured by $(1-HHI)$ and the portion of off-exchange trading, is associated with a 1.2% and 0.26% increase in ISO costs respectively. Considering that these two fragmentation measures are concerned with how disperse trading volume is, an alternative measure, focusing on the degree of inter-exchange disagreement on volatility is used as a further check. Disagreement in volatility is measured as a volume weighted standard deviation in volatility measures across exchanges. Using the prices reported from separate exchanges, each exchange's volatility is measured as the realized variation (RV) of its price series. RV is measured as the sum of squared first differences in log prices, $RV = \sum_t (\ln P_{t+1} - \ln P_t)^2$, with corrections for microstructure noise biases following Zhang *et al.* (2005). I use RV estimates because the vast amount of price data available allows for a high level of confidence in the estimates. The motivating logic behind this measure is the idea that, for a security traded on many exchanges, in a well integrated market both the NYSE and NASDAQ should agree on how volatile the stock price is. I found that a 1% rise in RV-dispersion is associated with a 0.18% rise in the ISO premium. The relationship between fragmentation and ISO premiums is very robust, it persists across different measures and methodologies, across large

and small cap stocks, with and without the COVID-19 episode, and after introducing Heckman selection bias controls.

Related Papers

In terms of theoretical background, Glosten (1994) found that in the presence of an open electronic order book, a multi-venued trading environment provides the same trade executions as-if there were only a single consolidated limit order book (CXLOB). A corollary to this result holds that any additional competition to the order book is either unprofitable or redundant. In this context, the recent proliferation of trading venues is somewhat puzzling. The apparent divergence from theory is due to the violation of Glosten's underlying assumptions; as a result of these violations, the Glosten (1994) result fails to hold and market fragmentation is made costly. Of the assumptions made by Glosten, two fail to hold in reality: (1) Investors can costlessly and simultaneously trade against separate order books on different exchanges, and (2) bid and ask quotes cannot be cancelled while a trade is being executed. As I previously mentioned, regular market orders can only execute one at a time against NBBO quotes, this violates the first assumption. The second assumption is violated because the quotes on other exchanges may be revised while the MO-trader waits for the NBBO to update. ISOs are split trades that can simultaneously execute on multiple exchanges at once, and this simultaneity means that resting orders can not react mid-trade. Glosten's assumptions hold for sweeps, but only when the scope of execution is limited to that subset of sweepable quotes the ISO aims to interact with. The subset of total liquidity an ISO implementation aims to interact with is equal to the set of protected quotes in addition to the set of protected quotes in addition to the resting orders on the exchange it's trading through to. The ISO fill is equal to that of a similar trade executed in a hypothetical would where the total liquidity

across venues equalled the ISO subset in our world and Glosten's assumptions held.

More broadly, there exists a long history of theoretical papers detailing the potential pros and cons of a fragmented market. Madhavan (1995) and Pagano (1989) highlight the positive network externalities of a single exchange, arguing that "liquidity begets more liquidity" and that splitting liquidity across multiple exchanges squashes this positive effect. Chowdhry and Nanda (1996) posit that adverse selection is greater in a multi-venue environment. On the other side Economides (1996) argues that the monopoly overhang associated with a single trading venue outweighs any benefits from consolidation. Harris (1993) takes a different tact, saying that different exchanges with different trading rules may attract different traders, expanding the total trader base and hence improving liquidity. Along these lines the Boehmer and Boehmer (2003) empirical study documented an improvement in liquidity in the SPY once the NYSE started trading it on their exchanges.

This paper also fits along side a series of post Reg NMS empirical studies focusing on the effects of fragemented markets. O'Hara and Ye (2011) compare the execution quality of stocks with more and less exchange-dispersed trading volume and find that more disperse trading either improves or at least has no effect on market quality. In a more recent paper Haslag and Ringgenberg (2020), similarly find that market quality improved as markets "fragmented", although the improvements accrued mostly to large stocks. As for measuring ISO costs, Chakravarty *et al.* (2012) find that ISO trading is more informed and have higher effective spreads than non-ISO trades, though they do not address the issue of fragmentation.

In this paper I identify the mechanical differences between sweep and non-sweep orders as a channel through which fragmentation effects modern markets. I show that traders make use of ISOs in order to avoid the delays and price uncertainty of the fragmented environment. I contribute to the market fragmentation literature by

developing a methodology to measure the premium paid to avoid fragmentation and putting a dollar value to these costs. Using new and old measures of fragmentation, I found a positive relationship between ISO premiums and market fragmentation. Given that ISOs are used to execute nearly half of on-exchange trades, these trading costs are of real practical importance and deserving of regulatory consideration, especially if changes to trading rules would substantially effect how sweeps operate.

The rest of the paper is organized as follows: Section 2 describes the current trading landscape, the regulatory environment and the mechanics of ISOs along with an illustrative example. Section 3 covers the methodology and data used to measure ISO costs; section 4 does the same for market fragmentation. Section 5 covers the empirical design and results while section 6 concludes.

1.2 The Trading Landscape

Regulation National Market System

Modern US markets were largely shaped by the SEC's Regulation National Market System (Reg NMS). Reg NMS is a set of SEC rules first introduced in 2005 and implemented in 2007. Reg NMS had sought to integrate U.S. security exchanges by setting common rules by which exchanges may carry-out trades and post quotes. The rules of consequence here are Rules 602, 604, and 611.

Rules 602 and 604

Rules 602 and 604 requires exchanges to immediately post and update their best-priced visible quotes onto a consolidated tape. These quotes are known as "protected quotes" and represent firm commitments of price and quantity at which the exchange will honor incoming trades. Hidden orders are not included in protected quotes,

neither are orders for less than 100 shares. Protected quotes change or update in response to an order cancellation, a new posting, or the fulfilment of a trade. The consolidated tape association (CTA) and security information processors (SIPs) consolidate the updates from different trading venues. The daily millisecond TAQ data used in this paper provides a record of all protected quote updates put together from the CTA and SIPs.

Rule 611

Next, Rule 611, colloquially known as "the trade-through rule", requires that all trades be executed at the NBBO prices or better. Per the regulation, the NBBO is defined as the highest priced bid and lowest priced ask from the existing protected quotes on the tape. This rule requires that orders be routed to the exchanges with the best protected quotes; exchanges are not allowed to fill any order at a noncompetitive price. The onus of compliance lies on the trading venue. As a result, an exchange will either cancel back or reroute a market order to another exchange if they can not match or beat the NBBO, in order to ensure compliance. The intention of these rules was to resolve potential conflicts of interest in order-routing and enforce a common price-priority in trade executions across public trading venues. If resting limit orders remain in place as the MOs, which make up a trade, execute across exchanges, the trade would get the best execution possible given the entirety of the posted liquidity.

Intermarket Sweep Orders

Trading against all the displayed liquidity across all exchanges requires a fragmented trading strategy with a lot of starts and stops permitting ample time for the market to move against the trader. The Reg NMS rules have unintentionally allowed market fragmentation to affect normal trading by depriving traders of the

instantaneous book climbing characteristic of a single-venue market structure. There is however an alternative to trading in this manner.

The way around this is the ISO which is Rule 611 exempt. ISOs have obtained widespread use, roughly half of all sample trades examined here were executed using an ISO. Since they are Rule 611 exempt, an exchange is free to fill an ISO even if they're not quoting at the NBBO. When sending an ISO to a particular exchange, the sender commits themselves to sending concurrent ISO orders to clear out any protected liquidity posted at better prices on other exchanges. To be protected, a quotation must be the "best bid" or "best offer" of a national securities exchange.[5] Since ISO senders commit themselves to clearing out all more competitively priced protected volume, the ISO exception appears to be aligned with the general spirit of Rule 611.

Example

Consider the case where a security is traded across the three exchanges A, B, and C. The prevailing protected and un-protected quotes for the exchanges are displayed in Figure 3. The bid and ask quotes of each exchange's order book lie on the top and bottom respectively with the most competitive quotes at the top. The blue cells represent that exchange's protected quotes and the green cells correspond to unprotected liquidity. Hidden orders are included in the grey font for illustrative purposes; note that they do not qualify as a protected quote. Across the three pairs of protected quotes, the best bid (100 @ 5) and best ask (50 @ 6) (in bold) make up the NBBO.

A trader looking to sell 200 shares in this environment may decide to do so by

[5]See the April 2015 "Rule 611 of Regulation NMS" memorandum by the Division of Trading and Markets of the U.S. SEC: https://www.sec.gov/spotlight/emsac/memo-rule-611-regulation-nms.pdf

piece-wise submitting smaller market orders submitted sequentially or by the simultaneous submission of ISOs. The trade, using ISOs, is made up of three sell ISOs of 100 to exchange A, 50 to exchange B, and 50 to exchange C. Executed at the prevailing prices, the sale yields proceeds of $950. Assuming that there is no price-movement in-between market order submissions, a trader would first submit a buy order for 100 to exchange A, wait for the NBB to update to the exchange B before clearing that out with an order for 50, and finally send another order of 50 to exchange A which would then be the prevailing NBB at a price of $4.50. Trading in this manner would yield proceeds of $975. On its face it appears as the ISOs are dominated, since the MOs had access to better priced unprotected quotes, but it is possible that the market prices change mid-trade when using MOs. To illustrate this suppose that the orders across the exchanges are cancelled-back $0.50 to the prices in red after the first market order for 100 is filled. Were this to happen the MO-trading would instead yield $925. Non-ISO trading promises the best prices in a perfect world but an uncertain fill in an imperfect one whereas ISOs are quicker and more sure, though possibly at worse prices.

Market Fragmentation

The ISO-MO dichotomy can be expressed in terms of integrated and fragmented markets. Glosten (1994) shows that there should not be a difference in execution quality between a single or multi-venued market. His result rests on the assumption that trades can be costlessly split into multiple orders and simultaneously execute against multiple exchanges. This is not true for MO implementations, making fragmentation costly, but is true for ISOs, and hence trade as-if there were a single LOB. In terms of the example, you can see this by first considering a hypothetical LOB constructed from the prevailing quotes (Figure 4). This ISO implemented trade has

the same execution as a MO would if the hypothetical book was the only game in town. In contrast, now imagine a "combined book" which includes all protected and unprotected quotes in the different exchanges. The sequential MO implementation of the trade mimics that of a MO against the combined book, though with breaks in time between orders during which the book can shift mid-trade. Quotes may randomly move mid-trade, but, more importantly, the changes may be strategic. For example, a vigilant liquidity provider could rationally expect the arrival of additional sell orders after observing a MO clear out the available NBB liquidity on one exchange. In response to this first MO, the liquidity provider may revise their resting buy limit orders on other exchanges down to a more advantageous price to take advantage of the predicted order flow.

It is worth highlighting that what is missing from the MO implementation is the instantaneity which is characteristic of an integrated market. Differences between the executions of ISO and non-ISO trades reflect a sort of "price of immediacy". I hasten to emphasize that it is the lack of time precluding quote revisions which is of first-order importance, and it is not the actual milliseconds saved that matters here. If quotes could not be revised mid-trade, trades would have the same execution whether it took 1 second or 0.1 seconds to complete.

Since the MOs trade against a "fuller" book, executing the trade with non-Rule 611 orders is always at least as good as the ISO implementation so long as the book does not change. If however, the market is such that small changes in the quotes may have an out-sized impact on execution quality, an ISO may offer a more certain fill. Simply put, ISOs offer an immediate execution with little price uncertainty whereas MOs offer a slower fill at a potentially better but uncertain average price. A priori, it is not clear if ISO traders should be compensated for trading against an incomplete book or if they should be paying a premium to trade in a pseudo-integrated environment.

1.3 Measuring ISO Costs

I measure the costs of market fragmentation by comparing the executions of ISO and non-ISO trades. ISOs trade against a subset of the available liquidity as-if the markets were integrated whereas non-ISO trades are effected by fragmentation. Being able to effectively measure differences in ISO and non-ISO trades is paramount. In this section I outline the data and different methods I use to construct measures of this difference.

Trade Excess Costs

I define a trade's excess cost (TEC) as the dollar amount that could've been saved (or lost) had the trade been executed against the prevailing NBBO quote first. For Rule-611 exempt orders, such as ISOs, the TEC proxies the difference in transaction costs with a counterfactual MO implementation. Recall the first example of an ISO purchase for 100 shares clearing at \$5 per share even though there were 50 shares available at the NBO of \$4. Under the ideal MO implementation, 50 shares would first be purchased at \$4 and the next 50 shares purchased at \$5. By "ideal" I mean that quotes are assumed to not move mid-trade under the MO counterfactual, so that the implementation is possible. In the example, the TEC for the trade is \$50 or \$0.50 per share. It is worth noting that the TEC may be negative if the order received price improvement and transacted at a price superior to the NBBO. Formally, for an order of size Q and transaction price P, the TEC is calculated as:

$$TEC = \begin{cases} Q \times (P - NBO) \times \text{Coverage}(Q; NBO) & \text{for buys} \\ Q \times (NBB - P) \times \text{Coverage}(Q; NBB) & \text{for sales} \end{cases} \tag{1.1}$$

Where the trade coverage is the proportion of Q which could be absorbed by the

NBBO quote of size Q_{NBBO}:

$$\text{Coverage}(Q;k) = \text{Min}\Big[1, \frac{Q_k}{Q}\Big] \quad \text{for} \quad k \in \{NBB, NBO\} \tag{1.2}$$

Coverage equals 100% if the liquidity available at the appropriate NBBO quote is greater than the size of the trade. Taking coverage into consideration is important so to only count these shares which could've plausibly been transacted at the NBBO. If a trade is for 1,000 shares but only 100 are available at the best quote, treating the trade as if 100% of it could be traded at the better price (even though only 10% could be) would overestimate the total costs by a factor of 10.

To frame it somewhat differently, the TEC corresponds to the difference in effective spreads between two trades scaled by the trade coverage. Typically the effective spread, the distance from the NBBO midpoint ($P-M$), is used as a trade's transaction cost. Instead of the midpoint the TEC tracks the distance between the transaction price and the NBBO quote ($P-NBO$); this corresponds to the difference in effective spreads between the actual trade and a trade which executed at the NBBO, i.e. $P-NBO = (P-M)-(NBO-M)$. The TEC is also measured in basis points after scaling the per-share TEC by the price, $TEC_{bps} = \frac{TEC}{P \cdot Q}$.

The main benefit of the TEC measure is its capacity to proxy the difference in execution between the ISO and MO alternatives without having to observe the unprotected quotes. As explained in Section 1.2, the comparison between an ISO and MO executed amounted to a straight forward comparison between the hypothetical and combined books. However, since unprotected quotes are not included in the TAQ data used in this paper, such a direct comparison can not be made. In contrast, only trades and the prevailing NBBO quotes (which are reported in the TAQ data) are needed to calculate a trade's TEC.

Going back to the ISO example of Section 1.2, the prevailing NBBO could be

derived after observing the first three quote updates in Figure 5. The first trade order observed is an ISO sale of 100 shares on exchange A; since the order was executed at the NBB it had a TEC of 0. In response to the first ISO, exchange A updates its protected quote to 100 shares at \$4.5. The second ISO to sell 50 shares on exchange B also has a TEX of 0; after this trade, the NBB is updated to the quote on exchange A at \$4.5. The last ISO observed in the data is a sale of 50 shares at \$4. This last ISO has a TEC of \$25 because 50 shares were sold at \$4 despite there being 100 shares available at \$4.5.

Despite the benefits of using the TEC, it is a noisy measure. The source of this noise is a sensitivity to the order in which the trades are reported in the data. If instead, the exchange C ISO was reported first, the prevailing NBB would've been \$5 rather than \$4.5 and the TEC would be calculated as 50 rather than 25. An ISO can be filled by an exchange without having to wait for the NBBO to update; it is possible for the execution of one ISO to be reported onto the tape before the quote from another exchange is updated after filling the other ISO.

Gross and Relative Daily Aggregates

The TEC is a measure of the whole dollar amount of benefit from executing a single order, at least in part, against the NBBO quote rather than where it actually transacted. TEC is a "per-order" measure and not necessarily a "per-trade" measure; an agent implementing a trade via five ISOs produces five TECs and not just one. All the individual orders would be marked as an ISO, the ones which take out the protected quotes as well as the trade-through orders. When an ISO-implemented trade "walks the book" on a given exchange, it is reported as a series of trades at the different prices.

Measuring the TEC on a per-trade measure requires deciding which orders corre-

spond to one trade and which ones correspond to another, a difficult task. Instead of grouping ISOs together into individual trades, I aggregate the ISO TECs up to the daily level for each security to obtain a gross ISO measure. A daily per-share ISO cost measure is calculated as the gross measure divided by the total ISO volume for the day:

$$ISOgross = \sum_{i \in ISO} TEC_i \;\; ; \;\; ISOcps = \sum_{i \in ISO} \frac{TEC_i}{Q_i} = \sum_{i \in ISO} TECcps_i \tag{1.3}$$

The aggregated basis point measure is simply a volume weighted average of the TEC bps for ISOs. Aggregating trades across the span of the entire day has the additional benefit of dampening the potential noise in the TEC measure coming from the order sensitivity. If the ISO orderings in the data are random, the resulting averages would be unbiased.

Relative ISO Costs Per Share

The aggregate TEC measure is a rough proxy for the average difference between an ISO implementation and the perfect MO implementation. In reality MO implementations are not perfect and the actual trade-off is between an ISO implementation and an imperfect MO implementation. Non-ISO trades may carry non-zero TEC measures if they interact with hidden liquidity, are odd-lot portions (< 100 shares), is reported out of sequence, or if there are delays/noise in the updating of the NBBO. To this end, I measure the relative difference in TEC measures between ISO and non-ISO trades.

$$RISOcps = \sum_{i \in ISO} \frac{TEC_i}{Q_i} - \sum_{i \notin ISO} \frac{TEC_i}{Q_i} \tag{1.4}$$

The setup here is similar to that employed in a difference in differences; the ISO designated trades comprise the treatment group and the non-ISO trades are the control group.

Taking the relative difference addresses any bias inherent to the TEC methodology as well as taking into account the possibilities of actual non-ISO orders realizing price improvement. The reasoning here is that if the TEC methodology is systematically biased one way or another it would also manifest in non-ISO trades. If such a bias exists, subtracting the non-ISO from the ISO costs should wash it out.

Alternative Methods

I compare trades which occur in the same 15 minute time interval as well as trades which occur right next to each other to minimize the effect of intraday variation in market conditions. If the choice between an ISO implementation and the MO alternative is largely driven by intraday market conditions, going up to the daily level may be too coarse of an aggregation. The concern is that the comparison of aggregate measures may reflect differences in market conditions rather than a difference in execution preference. To illustrate, suppose ISOs are more likely to be used when market depth is high and MOs are preferred if depth is low. If market depth varies substantially within each trading day, ISOs and MOs will be clustered at different times, so the ISO premium may just be the difference between deep and shallow markets.

Comparing Neighbor-Trades

One way of trying to address the aforementioned concern is to directly compare the ISO and MO trades closest in time proximity. The assumption is that market conditions are much more likely to be similar for trades which occur right next to

each other than for trades throughout the day. In order to do this comparison I develop a simple algorithm to group individual sweep orders into the same trade. Consecutive sweep orders are grouped into the some trade if: 1) the orders are in the same direction, 2) are executed on different exchanges, and 3) is within one second away from the first order in the sequence. A consequence of this relatively conservative filter is that not all ISOs will be grouped into trades; ISO trades which take more than one second to execute or have orders from the other side inter-weaved in their reporting would be cut short. Additionally, only those ISO trades made up of multiple ISOs are considered.

After grouping the individual orders into separate ISO and non-ISO trades,I next group adjacent ISO/non-ISO trades into pairs. In order to not double count any trade block, only those ISO trades with a non-ISO block immediately preceding it are counted in the analysis. For each i^{th} trade pairing, the relative ISO costs per share is simply the difference between the ISO and non-ISO TECs per share:

$$NRISOcps_{(i)} = \frac{TEC_{iso,i}}{Q_{iso,i}} - \frac{TEC_{niso,i}}{Q_{niso,i}} \tag{1.5}$$

A daily measure for the ISO premium is computed as volume-weighted average of NRISO costs for the day:

$$NRISO = \sum_i w_i NRISO_i \quad ; \quad w_i = \frac{Q_{iso,i}}{\sum_j Q_{iso,j}} \tag{1.6}$$

Figure 8 provides a graphical illustration of how this procedure compares to the daily aggregation,

15 Minute Slices

One way to avoid the concerns of daily aggregation is to aggregate to a finer interval. To this end, I also split each trading day into 15 minute slices and instead use the

15 minute time interval as the unit of time for the analysis. Any concern of varying market conditions muddying the relative ISO measure is addressed with the finer 15-minute time interval insofar that market conditions remain relatively constant within a 15 minute span as compared to the whole day.

Data and Sample Selection

In this study I use high frequency millisecond daily TAQ data and consider a selection of 2,933 securities for the sample period spanning January 1, 2019 to April 30, 2021. Security selection begins with the set of (CUSIP, Trade Symbol) pairings which appear in the Daily TAQ Master Files on both the first and last trading day of the sample period. I only keep those securities which could be matched, on the basis of their CUSIP, with the CRSP daily stock file December 2018. A subset of 2,346 securities have share codes {10,11}, exchange code {1,2,3}, prices >\$1.00 per share, and a market capitalization greater than \$100 million; these securities make up 47% of traded dollar volume in December 2018. The remaining 647 securities, amounting to 45% of trade volume, are comprised of 195 foreign securities, 89 REITs, 355 ETFs, and 13 other common stocks. All together, the complete sample[6] makes up 92% of the total traded volume in December 2018.

The total value of the traded volume is highly concentrated amongst a relatively small set of securities. The top 50 highest traded securities by dollar volume make up 39% of the total dollar volume listed in US public markets on December 2018. The median number of trading venues for these 50 securities is 8, which is greater than the overall median of 6. Summary statistics for different security sub-samples are reported in Table 1.

[6]The final selection is comprised of the union of the smallest set of securities needed to reach 90% of dollar volume and the 2,346 securities resulting from the aforementioned filtration.

Larger securities are more likely to be traded across more venues; the median number of public exchanges for the smallest 20% securities by market cap is 2 versus 8 for the largest securities. In absolute terms the relative ISO costs per share is lower for larger securities but this seems to be driven by the fact that transaction costs as a whole are smaller for these securities. When scaling the RISO cps by the average effective spread, the ISO costs are larger for the big securities. This pattern continues when breaking up securities by the number of trading venues with the scaled RISO cps monotonically increasing in the number of venues; the size of the securities also rise with the median number of trading venues. Of interest is that, after controlling for the size of the typical transaction costs, ISOs carry a larger premium when a security is traded across more venues.

Throughout the analysis what is counted as NISO volume are those regular trade orders identifiable in TAQ data as being without any special designation (other than "I" for odd lots) for it's sale condition. The logic is that these regular trades represent standard market orders and are the most suitable substitutes for an ISO trade. Unlike contingent, acquisition, block, cross, and same-day settlement cash trades ISOs do not require pre-negotiation with counter-parties. By their nature, the TEC of late reported or out of sequence trades would be mispecified and are therefore excluded; similar concerns preclude the inclusion of off-exchange trades. Also excluded from NISO classification are opening/closing prints, after-hour, and derivatively prices trades.

It is common practice for exchanges to make available to agents, at a fee, subscriptions to direct feeds to the exchange which makes pertinent market data available fractions of a second before it's reflected on the consolidated tape. Feeds such as the NASDAQ BookViewer, the CBOE BookViewer for the BATS exchanges, Arca Book for the NYSE, and the IEX's DEEP are examples of such services. It is for this reason

that I use participant timestamps[7] are used to order market data.

1.4 Measuring Market Fragmentation

The argument thus far has been that differences in execution costs between ISO and non-ISO trades reflects the premium paid by traders hoping to trade with the immediate and certain trade executions characteristic a SLOB whereas MO implementations are vulnerable to the complications made possible by market fragmentation. Finding evidence of a positive relationship between the ISO premium and market fragmentation would further my interpretation of the ISO premium; but first, measures of market fragmentation are needed.

Previous Fragmentation Measures

A natural starting point when trying to measure market fragmentation would be to co-op the measures which have previously been used in the literature. Here I borrow extensively from Haslag and Ringgenberg (2020) as well as O'Hara and Ye (2011). Haslag and Ringgenberg (2020) measure trade fragmentation using one less a Herfindahl-Hirschman Index of trade volume across reporting venues.

$$(1 - HHI) = 1 - \sum_{exg} \Big(\frac{Volume_{exg}}{TotalVolume} \Big)^2 \tag{1.7}$$

When $(1 - HHI)$ is low, trade executions are concentrated amongst a few number of venues; when it is high, trading is more disperse. The reasoning behind using the Haslag and Ringgenberg (2020) measure is that, in a market where fragmentation is costly, more disperse trading exacerbates the costs of fragmentation. O'Hara and Ye (2011) primarily measure trade fragmentation using the fraction of off-exchange, or

[7]The time the participant venue made the message available to the SIP, rather than the time the message was published by the SIP.

"dark", trading $\frac{DarkVolume}{TotalVolume}$. Even if the public exchanges were effectively integrated, if the dark venues are not included in that integration, the market as a whole may still suffer from fragmentation. Using the fraction of off-exchange trade volume is akin to how the NASDAQ market share was used in studies like Bennett and Wei (2006). Over 30 non-public trade venues employ the FINRA Trade Reporting Facility (FINRA-TRF), identifiable in TAQ data with exchange code "D", to publish transaction details. Using these measures, I run *log-log* panel regressions of different measures of market quality on a constant and these two measures of fragmentation[8]. In line with their previous findings a negative relationship is found between these measures and the market quality variables, with this relationship being stronger with being stronger for the most heavily traded securities. The results from these regressions are reported in Table 2. It is worth noting that my sample is different in the cross-section, I do not restrict my selection to common shares, and include other security types such as REITs and ETFs. My sample also covers a different time period than the previous studies. Using the Haslag and Ringgenberg (2020) original measure (which includes off-exchange trades in the HHI), univariate regressions (unreported) of the form $y_{i,t} = \alpha + \beta \ln[(1 - HHI)] + \epsilon_{i,t}$ yielded coefficients consistent with their original results. These coefficients were of similar statistical significance, but of $\frac{1}{4}$ the magnitude of those reported in Table 2.

Of relevance here is that a positive relationship between the ISO costs and the fragmentation measures after controlling for security fixed effects. Gross, relative, and scaled relative costs all have positive and statistically signifigant relationships with the $(1 - HHI)$ and $\frac{DarkVol}{TotalVol}$ after controlling for security fixed effects. Broadly speaking, as market fragmentation increases, the ISO premium, regardless of how it's

[8]Haslag and Ringgenberg (2020) included off-exchange trades when computing the HHI whereas I do not.

measured, also rises. In line with the idea of ISO costs being driven by fragmentation, the relationship is much stronger, in terms of magnitude, for those securities which are tend to be traded across more venues.

Alternative Fragmentation Measure

The previous measures of fragmentation have largely been concerned with the degree to which trading volume is segmented, or split, across multiple trading venues. While it is doubtlessly true that the segmentation of trade across venues is an important part of fragmentation, however, unless that split trade results in measurable differences in trade executions, segmentation alone does not lead to fragmentation costs.

Having multiple trading venues is a necessary but not sufficient condition for fragmentation to exist. If the Glosten (1994) assumptions held and the market was well integrated, there would be no meaningful difference in trading as compared to a single-venued market. As the Haslag and Ringgenberg (2020) measure goes to zero, $(1 - HHI) \to 0$, the result is one exchange capturing 100% of trade volume. $(1 - HHI) = 0$ is a de-facto single venue market, and thus a sufficient condition for the absence of costless fragmentation but it is not a necessary one. For example, if Glosten's assumptions held, a market with a single CLOB would have the same executions as a market where the total liquidity was split across 20 identical order books. In the former case $(1 - HHI) = 0$ because there's a single venue, in the latter case, $(1 - HHI) = 0.95$; though both markets have the same executions. A high level of heterogeneity between dark venues characterize the off-exchange markets. On some dark venues all liquidity is hidden, while on other venues trading is not anonymized, these structures violate Glosten's other assumptions of quote visibility

and anonymity[9]. A large fraction of off-exchange trading is likely to be associated with costly fragmentation, though an absence of off-exchange trading does not preclude it.

In contrast to trade-segmented based fragmentation measures I instead focus on cross-exchange disagreement. As in Hasbrouck (1995) I assume, for each security, the existence of a "true" or "fundamental" latent price process. For the sake of mathematical completeness, suppose that this latent price process follows a generalized geometric Brownian Motion:

$$dP_t = \mu_t P_t dt + \sigma_t P_t dB_t \tag{1.8}$$

Each exchange throughout the trading day generates a series of noisy "observations" of this latent price, these observations may take the form of trade transaction prices, quoted prices, etc. With different price series', one from every exchange, the mathematical problem here is how to measure the extent to which these time-series agree/disagree with one another. I measure fragmentation along two dimensions: disagreements in spread, and disagreement in price volatility across exchanges. My reasoning for using RV dispersion is that since the RV serves as a measure for how volatile the latent price process is, then to the degree to which exchanges reflect the same information there should be no disagreement in their RV estimates. An exchange's over-spread is a rough measure of how much more it would cost to trade in that particular exchange relative to trading at the NBBO. The average difference in these spreads, the average over-spread across exchanges, captures the cost of taking liquidity exclusively from any particular exchange. To put it simply, in an integrated market the NYSE and the NASDAQ should agree on how volatile the price of AAPL is, and how much it should cost to trade it.

[9]The availability of real-time order book data, at a fee, and anonymous trading does not make this an issue for the public venues.

RV Dispersion

Why RV: Given the GBM assumption, the variance of the daily log returns would equate the quadratic variation (QV) of $d\ln P_t$ over the day. $\text{Var}[\ln P_{close} - \ln P_{open}] = \int_{open}^{close} \sigma_t^2 dt$. The realized variation (RV) of a price process is defined as the sum of squared first-differences in the log-price, $RV = \sum_t (\ln P_{t+1} - \ln P_t)^2$. Absent any noise in the price observations, P_t, the QV may be recovered exactly in the limit with the RV:

$$\lim_{n\to\infty} \underbrace{\sum_{i=1}^{n} (\ln P_{t_i} - \ln P_{t_{i-1}})^2}_{RV} = \underbrace{\int_{open}^{close} \sigma_t^2 dt}_{QV}, \quad \text{if } \lim_{n\to\infty} \sup_i [t_i - t_{i-1}] = 0 \tag{1.9}$$

Given the vast amount of intraday price data available in TAQ, The RVs, and by extension the volatility of daily returns, are estimated with immense precision. I employ a modification of the Zhang *et al.* (2005) methodology to allow for the full use of available data and correct for any microstructure noise contaminating the price observations (see Appendix A).

The dispersion in exchange RVs is measured as a trade volume-weighted standard deviation, formally:

$$Disp[RV_d] = \sqrt{\sum_E w_{E,d} (RV_{E,d} - \overline{RV_d})^2} \tag{1.10}$$

where

$$\overline{RV_d} = \sum_E w_{E,d} RV_{E,d} \quad \text{and} \quad w_{E,d} = \frac{tradeVolume_{E,d}}{\sum_K tradeVolume_{K,d}} \tag{1.11}$$

I calculated the RVs using transaction prices because prices, as intersections of supply and demand schedules, are only observed when trade occurs. As a robustness check I also calculated the RV using midpoint updates and using a five minute frequency for the subset of the 50 highest traded securities[10]. This exercise results in a Pearson

[10] I restricted this robustness exercise to only 50 securities dues to a substantially higher computational overhead associated with the use of quote rather than trade data.

correlation coefficient between the trade and midpoint-based RV estimates of 0.91 and a correlation of 0.87 between their first differences. The high correlation between RV first-differences suggests that estimates are similar throughout the time-series[11]; this can be visually checked in Figure 10 which also shows that the two measures are of similar magnitude.

Average Over-Spread

Another way of capturing a sense of fragmentation is to compare the quoted costs of trading across multiple exchanges at once with the cast of transacting at the NBBO. An exchange quoting a spread wider than the NBBO means that turning around a single share is more costly on that venue than at the NBBO. An exchange's over-spread at any point in time is simply how much greater its protected quote spread[12] is than the NBBO spread.

$$OS_{E,t} = Spread_{E,t} - \Big(NBO_t - NBB_t\Big) \tag{1.12}$$

When the over-spread is equal to zero, that exchanges posted quotes correspond to the NBBO; the less price competitive the quote is, the more positive the over-spread and the greater its "distance" is from the NBBO. In order to aggregate this measure of distance into a per-day quantity, I integrate the over-spread over the trading day. The average overspread (AOS) for the exchange is a time-weighted average of $OS_{E,t}$.

[11]The trade based RV averaged 5.7 bps with a standard deviation of 10 bps; the midpoint based RV averaged 6.7 bps with a standard deviation of 9.4 bps

[12]An exchange's spread is calculated post any taker fees and rebates:

$$Spread_{E,t} = (Ask_{E,t} - Bid_{E,t}) + r_E\%(Ask_{E,t} + Bid_{E,t})$$

where $r_E > 0$ for exchanges which charge takers of liquidity and $r_E < 0$ for exchanges paying a rebate to the takers of liquidity.

As before, the volume weighted AOS measure comprises the final AOS measure to be used in the analysis.

$$AOS_d = \sum_E w_{E,d} AOS_{E,d} \quad \text{and} \quad w_{E,d} = \frac{tradeVolume_{E,d}}{\sum_K tradeVolume_{K,d}} \tag{1.13}$$

An AOS of zero indicates that the cost of transacting a single share is the same across all the trading venues and is equal to transacting at the NBBO. A large AOS measure signifies that multiple exchanges are quoting spreads under than the NBBO for long periods of time.

Unlike previous measures of fragmentation which are determined by how order-flow is segmentation, these alternative fragmentation measures are not and provide an additional robustness test for the relationship between the premium and fragmentation. Using the RV dispersion, average overspread, and the NBBO midpoint RV instead of $\frac{DarkVol}{TotalVol}$ and $(1 - HHI)$ in the same panel regression exercises as before, a positive relationship is found between RV dispersion and AOS and measures of market quality. Unlike the previous measures, this relationship is robust to the inclusion fixed effects in the sample. The relationship between the dispersion in RVs and the ISO premium is stronger for securities which are traded across more venues vs securities traded on fewer venues. The NBBO RV is included to partial out the effect of volatility on $Disp[RV]$. I report these results in Table 3.

1.5 Empirical Design and Results

I employ panel regressions to check if there exists a positive relationship between ISO trade costs and various measures of market fragmentation.

Panel regressions of ISO trade costs on the various market fragmentation measures are used to check if there exists a positive relationship between the two. The existence of such a relationship would provide evidence for the contention that the measured

ISO costs are reflective of market fragmentation. My main results mainly focus on the relative ISO costs-per-share as the response as it is the most conservative of my measures. Specifically the regression specification is as follows:

$$y_{i,t} = \alpha + \beta_1' F_{i,t} + \beta_2' X_{i,t} + \epsilon_{i,t} \tag{1.14}$$

Where the vector of fragmentation measures is comprised of $F_{i,t} = \Big[\text{Disp}[RV]_{i,t}, AOS_{i,t}, (1-HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}\Big]$; with i, t indexing security and date.. The control variables captured in the design vector X includes $\{RV^{nbbo},\ Spread,\ sweepProp,\ ILLIQ\}$. The response variable y used in the main analysis is made up of relative ISO costs per share, though results using the non-relative measure of $ISOcps$ are reported in Appendix A. Both response and design variables are log-transformed to correct for skewness in level regression residuals; Figure 11 plots histograms demonstrating the difference in skewness in residuals. In order to control for economy-wide shocks over time and any unobserved security-level hetrogeneity month and security fixed effects are included. Of the control variables, RV^{nbbo} is a control for the securities volatility. Amihud (2002) $ILLIQ$ measure and the spread are used to as (il)liquidity proxies and the proportion of sweep volume is used to control for variation in costs due to any pre-existing proclivities for the preferencing of ISO trading.

The mainline results provide statistically significant evidence of a positive relationship between fragmentation measures and the sweeping costs. The most important fragmentation measures, in order of decsending magnitude, are $(1 - HHI)$, $\frac{DarkVol}{TradeVol}$, and $Disp[RV]$; a one percent rise in each of these fragmentation measures are associated with a 1.28%, 0.27%, and 0.18% rise in ISO costs respectively. While the other fragmentation measure carries a positive coefficient absent controls; the AOS coefficient switches signs and does not appear to be economically significant in its size. Despite the occasional importance of AOS, I will focus on the variables

$[(1-HHI), \frac{DarkVol}{TradeVol}, Disp[RV]]$, hereupon referred as the "main" fragmentation measures. The positive relationship between sweep costs and the main fragmentation measures is much more pronounced when looking only at the top 50 traded securities and stronger still when considering only those securities which are traded on more than 9 exchanges. The overall R^2 of the panel regression is substantially larger for the subsample of securities with many trade venues compared to the sample with relatively few trade venues. That the explanatory power of the fragmentation measures is greater for securities with more trade venues is consistent with the idea that the ISO premium is indeed driven by market fragmentation.

Sub-Sample Analysis

Large and Not-Large Securities

I next check to see if the results are robust to size, because large securities are more likely to be heavily traded across more exchanges than other securities and the evidence presented in Table 4 shows a particularly strong relationship for securities traded across more venues. This robustness check addresses the natural concern that the market dynamics in large/small securities may be substantially different than for other securities. As a way of investigating this issue, the sample of securities have been split into size categories of small, middle, and large market cap categories. The small category is comprised of the smallest 20% companies in the sample, large sized stocks are the largest 20%, and the remaining middle 60% comprise the mid-sized category. The positive relationship with the main fragmentation measures remains both economically and statistically significant across securities of different sizes. I estimate the panel regression for each size quintile to assess how the relationship between the cost to sweep and the fragmentation varies across securities of different

sizes. It does not seem to be the case that the relationship is substantially different across securities of different sizes since the main fragmentation measures remain positive and significant across the smallest to the largest securities. The importance of RV dispersion and the fraction of off-exchange trading remains relatively constant as the size of the security rises whilst $(1 - HHI)$ increases in magnitude with size. Explanatory power, as measured by the overall R^2 also increases with security size.

With and Without COVID-19

Given the time-span of the sample period, spanning from 1/1/2019 - 4/30/2021, the question arises whether the results are attributable to the unprecedented market conditions introduced with COVID-19. Observing a time-series of market volatility, a clear structural break is observable sometime in March 2020. It was during this month that, along with the first evidence of exponential viral spread in the United States, the first lockdowns and travel restrictions were announced. The data is thus split into two subsamples, a pre-COVID sample which runs from the beginning of the sample to 2/1/2019, one month prior to the market reaction, and a COVID-forward sample comprising of the rest of the sample. The regression results from using these two subsamples are reported in Table 6. In both subsamples the coefficients on the main fragmentation measures are positive and remain significant, statistically indistinguishable from the whole-period coefficients. This hold despite the standard errors of the coefficients naturally rising do to the lower number of observations.

Addressing Endogeneity

A potential source of bias may come from the fact that the decision of executing a trade with an ISO rather than using market orders is endogenously determined and not random. Throughout the trading day there may be periods during which market

conditions would make ISO trading more or less desirable than utilizing market orders. That we observe different levels of ISO activity during some periods more than others could introduce a selection bias. In the previous analysis, the data was aggregated up to the daily level as a means of mitigating the endogeniety problem; the argument being that by averaging over periods of more and less desirability, the results wouldn't be reflective of any underlying latent decision factors. To this end I under-take two additional robustness-test in which daily aggregates for the response variable are calculated using temporally close trades and a 2-stage Heckman analysis to control for selection bias.

Neighbor Trades

Table 7 reports the regression results for comparing neighboring trades. When compared to the baseline specification the positive relationship between the relative ISO costs per share and $Disp[RV]$ and $(1 - HHI)$ is lower when comparing neighboring trade-blocks. It is higher for the off-exchange trade fraction relationship. That the relationship is both greater in magnitude and with (marginally) smaller standard errors along with the notable increase in R^2 suggests that aggregating the trades up to daily level introduces noise rather than induce bias. Once again the positive relationship with the fraction of off-exchange persists here as well; even though the magnitude of the coefficients fall when comparing neighboring trades, the estimates are less volatile.

2-Stage Heckman Correction

Here the sample bias concerns are addressed more directly by implementing the Heckman (1979) two-stage estimation procedure which allows us to control for selection bias. Rather than aggregating up to the daily-level the data is split into fifteen minute

intervals. The first step is to estimate the propensity towards ISO trading via a probit model using possible drivers of the ISO decision.

$$ISOpercent_{i,t} = \Phi(Z_{i,t-1}\gamma + u_{i,t}) \tag{1.15}$$

By inverting the above equation a linear regression model may be estimated.

$$\Phi^{-1}(ISOpercent_{i,t}) = Z_{i,t}\gamma + u_{i,t} \tag{1.16}$$

The design variables Z are lagged by one period inorder to avoid a future bias and allow for a more causal interpretation. The choice of Z comprised of RV dispersion, volatility, the average quoted NBBO spread, and (1-HHI). These variables, along with entity and day effects, achieve overall R^2's greater than 50%. After estimating $\hat{\gamma}_i$, the inverse Mills ratio, denoted $\hat{\lambda}_{i,t}$ is

$$\hat{\lambda}_{i,t} = \frac{\phi(Z_{i,t}\hat{\gamma})}{\Phi(Z_{i,t}\hat{\gamma})} \tag{1.17}$$

Where $\phi(\cdot)$ is the standard normal probability density and $\Phi(\cdot)$ is the standard normal cumulative density function. In the second stage of the estimation the inverse Mills ratios are included in the regression of ISO costs on fragmentation measures as a control for sample selection.

$$ISOcosts_{i,t} = X_{i,t}\beta + \hat{\lambda}_{i,t}\theta + \epsilon_{i,t} \tag{1.18}$$

The test for sample selection boils down to whether or not the coefficients θ on the inverse Mills ratio are statistically significant or not.

The estimation results for the Heckman correction procedure are presented in Table 8. As before both response and design variables are log-transform in both stages of the estimation with the exceptions of $\Phi^{-1}(ISOpercent)$ and $\hat{\lambda}$. The exercise was conducted for both ISO costs per share and the relative ISO costs per share. In

both cases $\hat{\lambda}$ fails to rise above the 5% level of statistical significance which allows for the rejection of a sample selection problem.

The decision for whether or not to employ ISOs could be driven by unconsidered market conditions or some other third variable. Were this the case then we would observe more or less intense ISO trading due to variations in market conditions potentially presenting a kind of omitted variable bias in the regression results. Rather than having a daily sampling frequency, I measured trading over non-overlapping 15 minute intervals to test for selection bias. The first two columns of Table 8 reports the results of from the first-stage probit regression. Next. Table 8 reports the second-stage panel regression with the inverse Mills Ratio computed using the fitted values from the first-stage probit. Solely for the comparison purposes, I reestimate the 2nd-stage regression while excluding the inverse Mills Ratio and is reported in the final column. The coefficient on the inverse Mills Ratio is statistically indistinguishable from zero meaning that I can reject the hypothesis of a selection bias in the data.

1.6 Conclusion

Unlike a market with a single trading venue, a multi-venued market is susceptible to the effects of being fragmented. I take the view that markets are effectively more fragmented if outcomes are more sensitive to how trades are routed across venues, something dictated by the choice of order type. Taking the view that markets are fragmented in a multi-venued market when the manner/order in which a trade is executed across venues is of greater importance. Due to the nature of ISO executions, their trading costs relative to other trades should be reflective of the degree of market fragmentation. ISOs offer a method through which trades may be executed as-if the market were integrated. I found that traders are willing to tolerate executions at prices which are roughly 40% of the average effective spread worse than non-ISO

transactions. That traders routinely find it in their best interest to use sweeps, an instrument with a worse average execution price, to exempt their trade from market rules is informative to the ongoing policy debate regarding fragmentation. In this study I contend that this apparent premium is indeed driven by market fragmentation. I present evidence of a economically meaningful and statistically significant robust positive relationship between the ISO costs and a 1% increase in $Disp[RV]$, $(1-HHI)$, or $\frac{DarkVol}{TotalVol}$ are associated with a 1.28%, 0.29%, and 0.18% rise in relative ISO costs-per-share respectively. This positive relationship is very robust, it persists across different measures and methodologies, across large and not-large market cap sub-samples, with and without the COVID-19 episode, and after Heckman selection bias controls.

Chapter 2

THE PROFITABILITY OF LIQUIDITY PROVISION

2.1 Introduction

Continuous trading where investors can immediately execute buy or sell orders is made possible by the presence of counter-parties who stand ready to take on the opposite side of those trades. These collective counter-parties are said to provide liquidity to the markets by competitively supplying the quotes at which traders can buy or sell. Liquidity providers hope to buy low at the bid quotes to then exit the inventory position by selling at a higher ask price (and vice-versa), profiting from the spread between the two (Demsetz (1968)); on average liquidity providers do not realize the prevailing full quoted spread due to subsequent movements in the market quotes between trades (see, Kraus and Stoll, 1972; Hasbrouck, 1988; Stoll, 1989; Huang and Stoll, 1994). The provision of liquidity involves taking on risks associated with temporarily holding inventory such as adverse selection, price volatility, etc. In this paper, we measure the proceeds from the aggregate provision of liquidity and investigate the relationship between this aggregate realized profitability and the risk associated with providing said liquidity.

When measuring the realized profits from providing liquidity one has to match each inventory exacerbating trade to an off-setting trade where the inventory position is reversed, completing a "round-trip" trade. Absent the availability of trade-level data associated with individual liquidity providers, researchers have traditionally relied on proxies to gauge the returns to liquidity provision, the most important of which has been the realized spread. The realized spread rs corresponds to the signed differ-

ence between the transaction price P_t and the midpoint $M_{t+\tau}$ at some pre-specified horizon τ into the future.(Huang and Stoll, 1996; Bessembinder and Kaufman, 1997):

$$rs_{t,\tau} = \delta_t(P_t - M_{t+\tau}) \quad ; \quad \delta_t = \begin{cases} +1 & \text{if trade } t \text{ is buyer-initiated} \\ -1 & \text{if trade } t \text{ is seller-initiated.} \end{cases} \tag{2.1}$$

The realized-spread formulation is a mark-to-market estimate of profit, taking it as a literal measurement of the realized proceeds would assume that liquidity providers exit every trade-induced inventory position at the midpoint τ units of time into the future. The use of the realized spread measure has been so widespread that it was formally adopted by the SEC as a measure of market quality—Rule 11Ac1-5 (now Rule 605) requires market centers to disclose the volume-weighted realized spreads computed with a τ of 5 minutes. The reported Rule 605 data is often used by scholars seeking to understand the impact of market structure on trade execution quality.

The arbitrary choice of τ in the realized spread, which is left up to the researcher's discretion, represents a potential source of significant misspecification. The realized spread is a mark-to-market profit measured at a predetermined point in time and can substantially deviate from the realized proceeds if the price is different at the time of actual exit.[1] Furthermore, the amount of risk associated with each round-trip trade is directly related to the time it takes to complete the turnover. Longer waiting times increase the risk that the value of inventory held will decline, either due to random price changes or having been adversely selected by a better informed liquidity-demanding trader. In equilibrium, spreads would be competitively set by liquidity providers to compensate for the risk of bearing an inventory position (Glosten

[1]The importance of choosing the horizon at which to measure realized spreads has long been recognized by Huang and Stoll (1996): *"... If the period is too short, the subsequent price may reflect not a reversal but another trade in a series of trades pursuant to the same order. If the period is too long, unnecessary variability will enter into the measure..."*

and Milgrom, 1985). Employing a measure with a uniform τ for every trade can not, by construction, capture any of the variation in realized profitability due to heterogeneous inventory turn-around time. Even if a "sensible" choice of τ is used, if trades are reversed at various horizons the conventional measure of realized spread—using a fixed horizon for all trades—can deviate significantly from the true profits. Virtu, a prolific US market-maker, for example, reported negative average realized spreads (measured over a five-minute horizon under Rule 605) for 11 consecutive months during the calendar year 2019, despite their actual market-making profits being positive.

In contrast to the realized spread, we measure the realized profits to liquidity provision by directly tracking the round trips completed by passive liquidity providers in the aggregate. We take the view that each trade has a passive (liquidity providing) and an aggressive (liquidity demanding) side. Using existing technologies (Holden and Jacobsen, 2014) to identify the passive side of every trade, we track the aggregate inventory position as if a single "Aggregate Liquidity Provider" (ALP) supplied the liquidity to every trade. The ALP represents the aggregate provision of liquidity by the traders who take the opposite side of every liquidity-taking trade.[2] In effect, we are using limit orders as a proxy for liquidity provision. The realized profits of a round trip initiated at P_t and completed at $P_{t+\tau}$ is measured as:

$$rp_{t,\tau} = \delta_t(P_t - P_{t+\tau}) \quad ; \quad \delta_t = \begin{cases} +1 & \text{if trade } t \text{ is buyer-initiated} \\ -1 & \text{if trade } t \text{ is seller-initiated.} \end{cases} \tag{2.2}$$

In our formulation, we do not determine τ ourselves but rather every trade's τ is individually determined by an inventory tracking system and the presentation of the data. Our focus on the aggregate provision does not require the use of trader-labeled transaction data.

[2]The ALP takes on a positive inventory position when investors are selling, and a negative position when there's a preponderance of buyer-initiated trades.

We track the ALP's inventory position because we do not have data on individual liquidity providers. This means our measure could be contaminated by the inclusion of trades resulting from passive limit orders submitted by long-term investors who intend to acquire or dispose of a position (Foucault *et al.*, 2005). Despite this imperfection, we show that our measure does a better job at matching market-making revenues as compared to the five-minute realized spreads reported under the SEC's Rule 605. To illustrate, Figure 12 plots the volume-weighted monthly averages of the realized spreads reported by Virtu under Rule 605, and our measure of realized profitability. The two measures are plotted against a backdrop of Virtu's market-making revenue (from their quarterly and annual SEC filings) from September 2018 to January 2021. In contrast to the self-reported 5-minute realized spreads, which bear little relation to the general trend of trading revenues, our realized profitability, despite applying to liquidity provision in aggregate, much better captures the broad pattern of market-making revenues of Virtu.

The key feature that distinguishes our realized profitability from the conventional realized spread measure is the determination of the trade turnaround time τ, which requires us to match each trade with a subsequent offsetting trade (to form a round trip). To match offsetting trades we adopt a LIFO (Last-in First-out) inventory tracking system under which offsetting trades are matched with the most recent positions of the ALP, consistent with the fact that liquidity providers prefer a quick turnaround.[3] Our reliance on a set inventory tracking system essentially allows the data to determine τ as opposed to the researchers' arbitrary choice. Note that under any inventory tracking system, not all trades will be matched with an offsetting counterpart on the same day; we restrict our analysis to trades that are turned around

[3]We report results and discuss the methodological differences of using alternative inventory systems such as FIFO (First-in First-out) in Appendix B.

within a day. This restriction is based on the rationale that liquidity providers very often do "go home flat" and that limit order executions not offset on the same day are more likely to be trades by longer-term investors (Easley *et al.*, 2011). Using a sample of all common stocks in the US equity market from 2017 to 2020, we are able to identify a total of 16.8 billion round trips.

Using this data, we document substantial variation in the horizon τ at which trades are turned around, and show that realized spreads, measured with a fixed τ for all trades, can deviate significantly from the realized profits to liquidity provision both in the cross-section and in the time series. To shed light on the causes and implications of these discrepancies, we first examine how realized profitability varies with the endogenous market-making horizon τ and compare that to the term structure of realized spreads documented in Conrad and Wahal (2020). We then show how the specification of common τ across all trades can cause systematic mismeasurement in the estimates of profits using realized spread and provide possible solutions.

Since longer inventory turnaround time typically implies a higher risk of market making—for example, higher probability of adverse information exposure and price volatility, the relation between τ and realized profits should reflect the risk-return trade-off faced by an average liquidity provider. We collect round trips into groups with similar turnaround times τ and compute the dollar-volume weighted average realized profitability for each group to construct a term structure of aggregate realized profitability similar to that in Conrad and Wahal (2020) to visualize the relationship between turn-around time and profitability. Conrad and Wahal (2020) measure realized spreads at varying prespecified horizons and document that the average realized spread decreases sharply with the time horizon τ used for the measurement. Contrary to the findings of both Conrad and Wahal (2020) and Hasbrouck and Sofianos

(1993),[4] we find aggregate realized profitability to be increasing in the market-making horizon. Specifically, it increases from 1.9 bps for quick turn-around round trips ($\tau < 1$ seconds), up to 6 bps for trips turned around between 9 and 10 minutes. This upward-sloping term structure is consistent with the risk-return trade-off faced by liquidity providers in a competitive market-making environment—when the expected turnaround time τ is large, the duration of inventory risk exposure is longer and, as a result, a higher return is required (by setting wider spreads).[5]

To be clear, what we do is calculate the average profitability only for those round-trip trades with a similar τ (for example, all trips with horizons between 9 and 10 seconds) and repeat for various values of τ to construct our term structure. We use the average value of the realized spread calculated using the same τ for every trade regardless as to whether or not the particular trades were actually turned around at that time when constructing the realized-spread term structure. To reconcile the differences in our results to those obtained in the prior literature we decomposed out realized profitability measure into a realized spread component measured with the endogenous τ and the effective spread at the exit $(t + \tau)$ of the round trip:

$$rp_{t,\tau} = rs_{t,\tau} + \delta_t(M_{t+\tau} - P_{t+\tau}), \tag{2.3}$$

where the τ is the horizon at which the inventory acquired at time t is turned around under our inventory tracking system. The differences with the constant τ realized spread term structure may come from heterogeneity in τ in the $rs_{t,\tau}$ component or from including the effective spread component $\delta_t(M_{t+\tau} - P_{t+\tau})$. We find that the average effective spreads, $\delta_t(M_{t+\tau} - P_{t+\tau})$, are relatively stable across horizons, so the

[4]Hasbrouck and Sofianos (1993) used spectral analysis on average mark-to-market proceeds of NYSE specialist inventory changes to infer a downward term structure in realized spread.

[5]The notion of being compensated for providing "immediacy" and then waiting to connect buyers and sellers extends back to Demsetz (1968).

differences in term structure primarily stem from the heterogeneity in τ across trades. In other words, the differences in term structure come from the selection of trades assigned to each τ rather than including every trade for every τ. We find that, once we use an inventory tracking system to determine the τ for each trade and only plot out the realized spread component the resulting term structure is still upward-sloping (rising from 0.2 bps for $\tau < 1$ seconds to 3.5 bps for $9 < \tau \leq 10$ minutes).

Next, we investigate how the average level and shape of the term structure in realized profitability differ for stocks that are expected to have a quick turnaround and stocks in which liquidity providers must hold on to their position for a relatively long time. This analysis serves two purposes, (1) it helps to understand how the τ-realized profitability trade-off in the cross-section (when variations in inventory turnaround time τ are well expected) differs from that in the time series (when variations in τ are less well expected); (2) it allows to study how the deviation of realized spread from realized profitability varies across stocks.

We sort stocks into quintile groups based on their average τ and construct the term structure for each quintile using the round trips of only those stocks in the group. We find that, in the cross-section, average realized profitability increases sharply across quintile groups: from 2.45 bps for stocks with the fastest turnaround (average $\tau = 56$ seconds) to 15.53 bps for stocks with the slowest turnaround (average $\tau = 213$ seconds). This is intuitive because market making in stocks with longer average inventory turnaround is expectedly riskier; when providing liquidity in a stock with a historically longer average turn-around time, competitive spreads should be set wider to compensate for the market-making risk. Consistent with this explanation, we find the cross-sectional difference is mainly driven by the effective spread component of the realized profitability in Equation (2.3), which increases from 1.39 bps to 14.36 bps. When looking at groups of stocks with similar turn-around times, the trade-off

between inventory turnaround time and realized profitability is drastically different for different stocks. Specifically, the term structure of realized profitability is sharply increasing only for the fastest group: from 1.2 bps for τ below 1 second to 7 bps for τ around 10 minutes. As for the slowest group, the term structure exhibits a downward slope (decreasing from 17.5 bps to 15 bps).

The differences in the shape of the within-group term structures suggest that the relevant risks/considerations faced by liquidity providers are qualitatively distinct across different stocks. Liquidity providers in the fastest group face intense competition in market making at extremely short horizons, they compete for the orders which are quickly turned around by posting quotes at more and more attractive prices, narrowing spreads. Market making at these horizons is much less risky for stocks with fast turnaround—competitive forces drive down the profitability commensurate with the level of risk at the fast end of the term structure relative to the slow end, resulting in the upward-sloping shape. In contrast, in the "slow" markets, the chances of a quick turnaround are lower because trades are more sparse—more elapsed time typically implies more volatility—and more likely to be informed. The downward-sloping term structure for these stocks paints a picture where the spread is initially set wide because inventory is rationally *expected* to take a long time to be turned around; the longer inventory is held, ex-post, the more likely it is that the market maker fell victim to adverse selection; however, when offsetting orders arrive unexpectedly quickly (only if simply by chance), a larger portion of the initial spread is captured.[6] We interpret the downward sloping term structure as suggestive that adverse selection is a greater issue for the competitive outcome in stocks with a slow

[6]The "unexpectedness" is reflected by the fact that, for stocks with a slow average inventory turnaround, the dollar volume at the extremely short horizons is very small compared to the total dollar volume.

expected turnaround,[7] consistent with Easley *et al.* (1996).

In contrast to realized profitability, the term structures of the conventional realized spread are similarly downward sloping for all groups (though for the fastest two groups, the term structures seem to suggest some reversal for horizons above one minute). The difference between realized spread and realized profitability decreases monotonically both in level and in the term structure as we move towards stocks with a slower expected inventory turnaround. Specifically, for stocks with the fastest expected turnaround, average realized profitability is 382% larger than realized spread even for the shortest horizon (within one second); the difference increases with the horizon—realized profitability is sharply increasing in τ whereas realized spread is largely decreasing in τ (from 0.40 bps for trades turned around within one second to 0.165 bps for τ between half and one minute before reverting to 0.23 bps for τ between 8 and 10 minutes). As for the slowest group, the difference between realized profitability and realized spread is much smaller: average realized profitability is 84% larger than realized spread for the shortest horizon; the difference increases with τ at a much slower rate as both term structures are decreasing in τ.

Because trades are turned around at variously different horizons, the above results suggest that mismeasurement in the estimates of profits using realized spread (with a common τ for all trades) can be large and also time-varying, especially for stocks with fast turnaround, of which the profitability is highly sensitive to the inventory turnaround τ. Indeed, Figure 13 shows aggregate realized spread (measured with 10s) is significantly lower than the realized profitability throughout our sample period with the difference spiking during periods with high market volatility (when variations in

[7]Note that we are not taking a stand as to whether the profits are too low or too high for any stock at any horizon because we do not observe the full cost structure of market making across varying horizons.

time to exit are likely large). Compared to fast stocks, the realized spread for slow stocks captures the dynamics of realized profitability relatively better, potentially due to the lower sensitivity of the profitability to τ. The fact that the realized spreads of both fast and slow stocks are much smaller than their realized profitability counterpart is driven by the effective spread on the exit trade which is not captured by realized spread. We show that adding the effective spread to the conventional realized spread not only brings it closer to realized profitability in levels but also in dynamics: the correlation between the two increases from 0.29 to 0.79 for fast stocks and 0.49 to 0.66 for slow stocks. We find that a fixed τ realized spread is less correlated with the realized spread component of realized profitability (0.59) than the average effective spread is (0.68); this suggests that the effective spread itself, which does not require any determinations of τ does a better job at capturing the time-series dynamics of the realized profitability than a misspecified conventional realized spread measure.

2.2 Realized Spreads and Realized Profitability

The Passive Liquidity Provider

We measure the liquidity provision profitability by tracking the trading profits of a hypothetical trader we call the passive aggregate liquidity provider (ALP), who takes the passive side of every trade. Absent the simultaneous arrival of perfectly off-setting aggressive market orders, every trade must have an aggressive (liquidity-taking) and passive (liquidity-providing) side. In the modern electronic order book markets of today, liquidity providers serve the role of market making by submitting limit orders on both sides of the book. Indeed, any trader who submits a limit order is, for that moment, helping to make the market. Our concept of the ALP is made up of all actors who, however temporarily, contribute to the provision of liquidity.

The ALP takes the passive side to every liquidity-demanding trade and is such the collective market maker.[8] By focusing on the passive liquidity providers as a whole (the ALP), our study aims to shed light on the profitability of the liquidity provision business as a whole.

However, as pointed out by Foucault *et al.* (2005), passive orders are not the exclusive province of dedicated liquidity providers, traders often use limit orders to take on long-term positions,[9] we refer to these traders as unintentional liquidity providers (ULPs). ULPs contribute to the cumulative inventory of the collective ALP by passively taking on their positions. If we want to interpret the realized profitability as a measure of profitability for intraday liquidity providers who go home flat, then ULPs represent a source of noise in our measure of profits to liquidity provision. In Section 2.3 we show how the usage of LIFO and robust tests using alternative inventory systems can alleviate such concern.

Realized Spreads as a Profitability Measure

Equilibrium bid-ask spread—quoted spread, effective spread—reflects both the costs of providing immediate trading (e.g., inventory holding, order processing, adverse selection) and competition between liquidity providers (e.g., Glosten and Milgrom, 1985; Stoll, 1978; Ho and Stoll, 1981; Ho and Stoll, 1983; Kyle, 1989). The empirical literature on the relation between bid-ask spreads and trade execution costs typically features a breakdown of the effective spread into a permanently component—price impact, measured as the drift in quote midpoint following a trade—reflecting

[8]The SEC defines market makers as firms that stand ready to buy and sell stock on a regular and continuous basis at a publicly quoted price.

[9]By "long-term" we mean that the trader intends to hold onto their position for more than a day, longer than the intraday market-making horizons targeted by liquidity providers that we study here.

the informativeness of a trade, and a transitory component, realized spread, reflecting the reversal in transaction price associated with liquidity provision (e.g., Glosten and Harris, 1988; Hasbrouck, 1988). By and large, realized spreads have been calculated as the drift of midpoint away from the trade price at some prechosen fixed horizon $\bar{\tau}$ in the future:

$$rs_{t,\bar{\tau}} = \delta_t(P_t - M_{t+\bar{\tau}}); \quad \delta_t = \begin{cases} +1 & \text{if trade } t \text{ is buyer-initiated} \\ -1 & \text{if trade } t \text{ is seller-initiated.} \end{cases} \tag{2.4}$$

This measure is typically interpreted as the residual profits captured by the liquidity providers following the realization of price impact from trades (from the decomposition of effective spread).

$$\underbrace{\delta_t(P_t - M_{t+\bar{\tau}})}_{\text{Realized spread } (rs_{t,\bar{\tau}})} = \underbrace{\delta_t(P_t - M_t)}_{\text{Effective spread } (es_t)} - \underbrace{\delta_t(M_t - M_{t+\bar{\tau}})}_{\text{Price impact } (pi_{t,\bar{\tau}})}. \tag{2.5}$$

Under the implicit assumption that the midpoint proxies for the fundamental value, what this signed difference (between P_t and $M_{t+\bar{\tau}}$) captures is a mark-to-market profit. A mark-to-market profit measure at one point can be way off as a measure of round-trip profit if the price subsequently moves before the actual sale.

Realized Profitability

In this paper, we seek to measure the proceeds from the round-trip trades, as opposed to mark-to-market estimates. We track the prices and quantities at which the ALP enters and exits inventory positions and compute the realized return of each round trip—we call this return "realized profitability." A round trip is a pair of (partial) trades that comprise a reversal in the ALP's inventory position. For instance, the ALP buying 10 shares from a seller in the morning and later selling 5 of those shares to a buyer in the evening would make a round trip for 5 shares. The proceeds of a round trip initiated by a time t trade at a price P_t and completed by

an offsetting time $t+\tau^*$ trade at $P_{t+\tau^*}$, where τ^* is the time horizon identified under LIFO, is defined as:

$$\text{RoundTripProceeds}_{t,t+\tau^*} = \delta_t \ |Q_{t,t+\tau^*}| \ (P_t - P_{t+\tau^*}), \tag{2.6}$$

where $\delta_t = 1$ if the initiating trade at time t was an aggressive buy and $\delta_t = -1$ if it's an aggressive sell and $|Q_{t,t+\tau^*}|$ is the number of shares reversed by the $t+\tau^*$ trade. The realized profitability $rp_{t,t+\tau^*}$ of the round trip is computed as the per-share return of the proceeds:

$$rp_{t,t+\tau^*} = \frac{\text{RoundTripProceeds}_{t,t+\tau^*}}{|Q_{t,t+\tau^*}|} = \delta_t(P_t - P_{t+\tau^*}). \tag{2.7}$$

In contrast to the realized spread (Equation (2.4)) which measures mark-to-market profits at a prespecified horizon $\bar{\tau}$, realized profitability measures the profits of a trader providing liquidity to both the initiating and reversing trades (using the τ^* at which trades are turned around).[10]

Similar to the interpretation of realized spread as a residual profit to liquidity providers in Equation (2.5), our realized profitability can also be interpreted as such a residual profit. Specifically, it is equal to the sum of the effective spreads at the initiation and termination of the round-trip trade less the price impact measured over the duration of the round trip as in Equation (2.8).

$$\underbrace{\delta_t(P_t - P_{t+\tau^*})}_{rp_{t,\tau^*}} = \underbrace{\delta_t(P_t - M_t)}_{es_t} + \underbrace{\delta_t(M_{t+\tau^*} - P_{t+\tau^*})}_{es_{t+\tau^*}} - \underbrace{\delta_t(M_{t+\tau^*} - M_t)}_{pi_{t,\tau^*}}. \tag{2.8}$$

Here the sum of the effective spreads at time t and $t+\tau^*$ reflect the full spread quoted by the liquidity provider for the round trip which is composed of both the entering and exiting trades.

[10]For example, if the ALP buys 1 share at the bid B_t and then sells that share later at the ask $A_{t+\tau^*}$ then the realized profitability would be $A_{t+\tau^*} - B_t$.

Substituting in the realized spread, $rs_{t,\tau^*} = \delta_t(P_t - M_t) - \delta_t(M_t - M_{t+\tau^*})$, into Equation 2.8 allows us to decompose the rp_{t,τ^*} into a realized spread component (with an endogenous τ^*) and the effective spread at the exit:

$$rp_{t,\tau^*} = rs_{t,\tau^*} - \delta_t(P_{t+\tau^*} - M_{t+\tau^*}) = rs_{t,\tau^*} + \delta_{t+\tau}(P_{t+\tau^*} - M_{t+\tau^*}) \tag{2.9}$$

Note that because the trade at time $t + \tau$ is an offset to the initial time t trade it's therefore the case that $\delta_{t+\tau} = -\delta_t$.

This decomposition helps illuminate any sources of differences between the realized profitability measure and the conventional realized spread on average. Starting with the simplified case where every LIFO determined turn-around horizon τ^* happens to be equal to the same constant $\bar{\tau}$, then the average realized profitability ($\sum_i(w_i \cdot rp_{t_i,\tau_i^*})$) would be equal to the average conventional realized spread with horizon $\bar{\tau}$ ($\sum_i(w_i \cdot rs_{t_i,\bar{\tau}})$) plus the average effective spread. After discounting the average effective spread (which is not effected by heterogeneity in τ^*), any difference between the realized profitability and fixed-τ realized spread in the averages would stem from heterogeneity in the LIFO determined τ^*s, $\sum_i(w_i \cdot (rs_{t_i,\tau_i^*} - rs_{t_i,\bar{\tau}}))$.

2.3 Methodology and Sample

Identify Round Trips

The main empirical challenge regarding the calculation of the realized profitability is how one decides which trades reverse one another to make a round trip. To construct round trips, we track the market-making inventory of the LP using trades of each stock. Specifically, for each stock, we record the LP's inventory entries starting from the first trade of a day: for example, a seller-initiated trade will count as the first positive inventory. Any following trades will be either recorded as a new inventory entry or used to offset the existing inventory entries depending on the sign

of the trade as compared to that of the existing inventory.

We primarily rely on a "Last In, First Out" (LIFO) inventory tracking system to decide which pieces of existing inventory are reversed by the incoming trades for two reasons: one, LIFO is economically appealing because it tends to match offsetting trades that are temporally closer (more likely from market makers), and two, everything else equal, an alternative system such as FIFO ("First-In, First-Out") introduces a mechanical bias in the estimates of realized profitability when there is large order imbalance.[11] However, for robustness, we also show that first, estimates of realized profitability are very similar under both LIFO and alternative tracking systems (FIFO and Weighted-Average-Cost) during days with small or no order imbalance, and second, for days with order imbalance, the general inferences from alternative tracking systems are the same as that from LIFO results when we properly control for the bias introduced by order imbalance.

Sample and Data Description

We use the daily Trade and Quote (TAQ) data from WRDS for the construction of round trips from January 5, 2017 to December 31, 2020. We use common filters on the CRSP universe for the selection of our sample stocks: all common shares (share codes 10 or 11) with exchange codes 1, 2, or 3. We also remove shares with a market capitalization below $100 million or a share price below $1 at the beginning of each year in our sample, to make sure micro-caps do not drive results. The CRSP sample is manually matched with the TAQ Masterfiles using the CUSIP code. We purposefully exclude trades that are likely to be missigned by the Lee and Ready algorithm, such as the opening prints (the first trades of the day) and trades reported late or out of

[11]The implementation of both inventory tracking systems, the comparison between the two, and the bias of the FIFO estimates during large order imbalance days are detailed in Appendix B.

sequence. We also drop block trades, orders designated with condition "B," or large trades with a size over the 95 percentile for trades for that stock, these kinds of trades are often prenegotiated and do not reflect the trades with which intraday liquidity providers typically interact with. Acquisition (A) and Cash Sale (C) designated trades are also dropped for similar concerns, even though such large trades are interesting by themselves, they are not the focus of this paper. For the trade signing, we use the quote and tick test from Lee and Ready (1991) following the implementation for daily TAQ data of Holden and Jacobsen (2014). A trader-initiated sell corresponds to an LP buy and a trader-initiated buy corresponds to an LP sale.

2.4 The Realized Profitability

Distribution of τ

We identified a total of 16.8 billion round trips. Figure 14 plots the distribution (histogram) of the turnaround time τ of all the round trips.

There is wide dispersion in τ across trades: although 79% of the volume has a turnaround time of fewer than 60 seconds; 8% has a turnaround time of more than 5 minutes. Importantly, when we decompose the dollar-weighted variance of τ into a cross-stock component and a within-stock component we find that nearly all of the variation, 97%, comes from the time series within each stock.

$$\underbrace{\sum_{i,t} w_{i,t}(\tau_{i,t}-\overline{\tau})^2}_{Total Variation} = \underbrace{\sum_{i} w_i(\bar{\tau}_i-\overline{\tau})^2}_{Across\ Stock} + \underbrace{\sum_{i,t} w_{i,t}(\tau_{i,t}-\overline{\tau}_i)^2}_{Within\ Stock} - 2\underbrace{\sum_{i,t} w_{i,t}(\tau_{i,t}-\overline{\tau}_i)(\overline{\tau}-\overline{\tau}_i)}_{Covariance}, \tag{2.10}$$

where $w_{i,t}$ is the dollar-volume weight for stock i's t^{th} trade. The fact that trades are turned around at variously different horizons even for the same stock suggests that, unless the profitability to liquidity provision is insensitive to the market-making horizon, selecting any fixed τ to approximate the profits with realized spread is unlikely

to be accurate. For instance, a τ of 60 seconds may be too short for some trades (e.g., large trades or partial trades from a large order)—short in the sense that price has yet to recover from the transitory drift caused by temporary order imbalance—but too long for other trades.

The discrepancy with realized spreads measured using a fixed horizon can be large especially during days with an abnormal amount of large or correlated orders (typically comes with high volatility in prices). To show this, in Figure 15 we plot the time series of the aggregate realized profitability together with the aggregate realized spread (at both 10 seconds and 6 minutes) and compare both time series with the realized revenue from market making reported by Virtu in their quarterly report. As one can observe, the realized spreads measured with both 10 seconds and 6-minute horizons fall far short of matching the time-series variation in Virtu's market-making revenue, especially during the highly volatile period in early 2020.

Aggregate Term Structure

To examine how realized profitability varies with the endogenous market-making horizon, we first sort all round trips into groups based on their turnaround τ (e.g., the first group contains round trips with τ between 0 and 1 second, the second group contains round trips with τ between 1 and 2 second, etc) and then for each group, we calculate the dollar-volume-weighted realized profitability (rp_τ). Such a structure allows easy comparison with the conventional realized spread, which is only defined at pre-specified horizons. Figure 16 plots the term structure of aggregate realized profitability, along with the corresponding effective spreads and price impacts from Equation (2.8).

We observe a clearly upward-sloping term structure of realized profitability which stands in stark contrast to the sharply downward-sloping term structure of realized

spread (Conrad and Wahal, 2020). The term structure not being flat along with a large amount of within-stock variation in τ means that any choice of a fixed τ in the calculation of realized spreads will lead to a misspecified estimate of realized profitability. We argue such an upward-sloping term structure is consistent with the risk-return trade-off faced by liquidity providers as a whole—slower inventory turnaround exposes liquidity providers to greater risk of, say, adverse information or large price swings; as compensation, they demand a higher return.

The accompanying term structure of effective spread reconfirms the above argument. As the turnaround time increases, effective spread also increases. This upward-sloping term structure of effective spread implies two things. First, market makers have rational expectations concerning the time it takes for a trade to be turned around. Second, they quote a higher spread for trades that they expect would take longer to offload—to compensate for the higher risk associated with holding the temporary inventory.

Sharpe Ratio Term Structure

The upward term structure of aggregate realized spreads provides a useful but imprecise depiction of the risk-return trade-off market makers face. To better visualize such a trade-off, we compute the Sharpe ratio (the ratio of dollar-volume-weighted average to the standard deviation of the realized spread) of all round trips in each τ group.[12] Figure 17 plots the term structure of Sharpe ratio.

In a perfect world absent frictions or costs, Sharpe ratios of liquidity provision

[12]For robustness, we also estimate the ratios in an alternative way: we first compute the Sharpe ratio of round trips in each τ group on a daily basis, and then compute a simple average of these daily estimates. The resulting Sharpe ratio estimates are almost the same using both methods.

across varying horizons should be equalized. In reality, however, frictions such as a high barrier to entry (e.g., high-frequency market making requires significant initial capital investment and operational costs) can limit competition thus causing deviation from equality. If we interpret the differences in Sharpe ratios across market-making horizons as reflecting such costs. The term structure in Figure 17 can also be viewed as the term structure of market-making cost. In Figure 17, market making at shorter horizons (within 1 second) exhibits a much higher Sharpe ratio at 7.3. This number declines sharply over the horizons up until 60 seconds and then slowly flattens out. Such a pattern is not surprising as marketing making at extremely short horizons is significantly more costly due to, say, data costs or server costs. At longer horizons above 5 minutes, we still observe an annualized Sharpe ratio as high as 2.8. By contrast, using the conventional measure of realized spread, the Sharpe ratio falls to almost zero after 1 minute. The evidence sheds light on the biases the conventional measure can generate, which we will discuss in more detail in the following section.

2.5 Dissecting the Term Structure

In this section, we break down the aggregate term structure and study both its cross-sectional and time-series components. To do that, we first compute the average τ for each stock using all round trips of that stock in our sample. We then sort firms into decile groups based on their average τ. With the grouping, we can separately study the time-series dimension of the term structure (within each group) and the cross-section dimension (across the groups).

Cross-sectional Variations in τ

The top panel of Figure 18 plots the distribution of stocks across varying τs. The y-axis denotes the percentage of stocks with average τ within the range marked by the

edges of the bars along the x-axis. The colors denote the decile groupings—the group with the fastest inventory turnaround is marked dark green whereas the group with the slowest inventory turnaround is marked dark red. The bottom panel of Figure 18 shows the (simple) average τ of stocks from each decile group.

As in Figure 18, the average inventory turnaround time is less than 200 seconds for more than 80% of all stocks. This is not surprising as we know the cross-sectional variation constitutes close to 0% to the aggregate variation in τ. The bottom decile group of stocks (the active group with the fastest inventory turnaround) has an average τ of 56 seconds. Whereas the average τ for the top decile group (the inactive group with the slowest turnaround) is 212 seconds.

Trade-off between τ and Realized Profitability in the Cross-section

Table 9 shows the dollar-volume-weighted average realized profitability for each decile group using all round trips of the stocks in that group. Realized profitability is strictly increasing in the average inventory turnaround time of a stock. The average realized profitability is 2.45 basis points for the stocks with the fastest turnaround time and increases to 15.53 for the stocks with the slowest inventory turnaround. Similarly, the term structure of exiting effective spread is also sharply upward sloping—increasing from 1.39 basis points to 14.36 basis points. The slope of this cross-sectional term structure is much steeper as compared to the aggregate term structure (raising from 3.2 to 4.6 for the same range in τ), reflecting a sharper risk-return trade-off in the cross-section: because the daily average turn-around time τ is relatively stable within a stock, liquidity providers should have a relatively good idea about the risk of market making in each stock and sets their quotes according to this perceived level of risk (increasing in τ). From Table 9 we see that the ALP is relatively good at pricing liquidity (setting the entering spread) in the cross-section and

gets compensated accordingly. This is consistent with Comerton-Forde *et al.* (2010) who find that market makers widen spreads as trading risks increase.

In terms of other characteristics of the stocks, Panel B of Table 9 shows that stocks with short turnaround times tend to be larger than those with longer turnaround times. They also have higher valuations (lower book-to-market ratios) as compared to slow stocks. This leads to the natural concern that the apparent relationship between τ and realized profitability is not driven by τ but rather other stock-level characteristics that just happen to be correlated with τ. To this end, we report in Table 10 the dollar-volume-weighted average rp for stock subsets sorted first by size and then average τ and also for stock subsets sorted first by book-to-market and then average τ. The initial sort serves as a rough means of controlling for size. The positive relationship between τ and rp remains intact for both small and large stocks. We repeat the same exercise with book-to-market and find the τ, rp trade-off to be similarly robust.

Trade-off between τ and Realized Profitability in the Time Series (within stock)

In this section, we investigate how the term structure of realized profitability differs for stocks whose trades are expected to be turned around quickly and stocks in which liquidity providers must hold on to their position for relatively longer. To do that, we construct the within term structure of realized profitability for each quintile group by estimating the dollar-volume-weighted average realized profitability at varying horizons using round trips of all stocks in that group. These within-term structures primarily reflect the τ-realized profitability trade-off in the time series. For those concerned with the cross-section variation within each group, we show that using an alternative estimation—compute the term structure for each stock (using dollar volume weights) and then aggregate all term structures by simple averaging

across all stocks within each group—yields almost the same results.

In Figure 19 we plot out, for all groups, the realized profitability term structure as well as the term structure of its components (the alternative realized spread and the effective spread at the exit). In contrast to the comparison of average realized profitability across the groups themselves, the relation between realized profitability and τ appears more complicated within each quintile group. Specifically, realized profitability is sharply increasing in τ only for those securities with the fastest inventory turnaround. The majority of trading, 83% (by dollar-volume), occurs in securities classified as "fast"; this causes the aggregate term structure to be upward-sloping. As we move towards the stocks with a slower turnaround time, the term structure begins to take on a downward slope (e.g., for the slowest two groups). This transformation from upward to downward sloping is even more pronounced when looking at the term structure of the realized spread component of the realized profitability. Similar variation in the term structure across securities does not emerge when looking at realized spreads. This is evidence of our realized profitability measure capturing aspects of the different markets which are missed by the conventional measure.

Figure 20 plots out the term structure of the realized spread, by using the same fixed τ for every trade, across the different groups of fast/slow stocks. We see a consistent downward-sloping term structure across the different groups with the only visible variation being in the gradient of the decline. The most important takeaway is that the fundamental trade-off between holding time τ and profitability is reversed for fast stocks. There are other implications as well. First, Huang and Stoll (1996) set forth the intuition that if the choice of τ is too short when computing realized spreads the observed price may not have reverted back to fundamental value, and if chosen too long it would be contaminated by the effects of other trades. By this logic, it should be the case that after a certain τ, the mean realized spreads should

level out as the additional noise is averaged out. We do not see this, for each group, we see realized spreads continue to decline even past the average turnaround time for each group; in fact, we see a (partial) reversal beginning to manifest in the two fastest groups (which together make up 95% of the whole market).

Figure 21 plots the average entering and exiting effective spread for the round trips at different turnaround times for the fast/slow groupings. In the graph, effective spreads are largely increasing in τ across groups, suggesting that the ALP quoted higher spreads for trades that took longer to turn around. If we take the (realized) time-to-exit as a reasonable proxy for the (expected) market-making risk, we can see the effective spreads increase with the expected risk of market making. We interpret that as evidence of the ALP's overall capability to evaluate the riskiness of trades in the time series, quoting a wider spread when trades take longer to turn around.

In terms of realized profitability, we attribute the differences in the term structures to the varying level of competition intensity across the groups. Specifically, for the group with the fastest inventory turnaround, market making at extremely short horizons is relatively less risky. This temptation of "risk-free" profits attracts intensive competition from market makers with speed advantages, driving down the profitability at these extremely short horizons. As we move towards stocks with a slower inventory turnaround time, the prospect of "risk-free" return gets slimmer as trades are sparser and more likely to be informative. For these stocks, concerns about information asymmetry and adverse selection discourage competition on quotes from high-speed market markers. As a result, the realized profitability is larger at the extremely short horizons and the remaining term structure is mostly dominated by price impact from adverse selection.

Term Structure Steepness and Volatility

Our interpretation of the term structure for both fast and slow stocks centers on a risk-return trade-off. One way to check this intuition is to see whether or not these trade-offs are more or less pronounced during periods of elevated price risk. Simply put, the *rp* term structure for fast securities should have a steeper upward slope when volatility is high and the *rp* for slow securities should be more downward sloping if during these times adverse selection risks are elevated. To measure the slope of the term structure we run monthly regressions of round trip realized profitability *rp* on the turn-around time τ and use the coefficient on τ as our measure of the slope. For fast stocks, this coefficient is positive indicating that the longer hold-times are associated with higher returns to the ALP on average, for slow stocks it is the reverse. We proxy for the level of risk by computing the realized variation of transaction prices calculated following the methodology laid out by Zhang *et al.* (2005). Figure 22 plots the slope of the term structures against the RV for both groups of stocks. We found that whenever the RV increases, the slope of the fast term structure becomes more positive while that of the slow group becomes more negative.

Deviation of Realized Spread From Realized Profitability

The previous section suggests that conventional realized spread measures can deviate significantly from our realized profitability. The difference between the two, however, is monotonically decreasing both in level and in term structure as we move towards stocks with a slower expected inventory turnaround. Specifically, for stocks with the fastest expected turnaround, the average realized profitability spread is 382% larger than realized spread even for the shortest horizon (within one second); the difference increases with the horizon—realized profitability is sharply increasing in τ

whereas realized spread is largely decreasing in τ (from 0.40 bps for trades turned around within one second to 0.165 bps for τ between half and one minute before reverting to 0.23 bps for τ between 8 and 10 minutes). As for the slowest group, the difference between realized profitability and realized spread is much smaller: average realized profitability is 84% larger than realized spread for the shortest horizon; the difference increases with τ by a much slower rate as both term structures are decreasing in τ.

Because trades are turned around at variously different horizons, the above results suggest that the biases in the estimates of profits using realized spread with a common τ for all trades can be large and also time-varying, especially for fast turnaround stocks of which the realized profitability is highly sensitive to the τ. Indeed, Figure 23 shows aggregate realized spreads (measured at both 10 seconds and 6 minutes) are significantly lower than the realized profitability throughout our sample period with the difference spiking during periods with high market volatility (when variations in time to exit are likely large). Compared to fast stocks, realized spreads for slow stocks capture the dynamics of realized profitability much better, potentially due to the lower sensitivity of the profitability to τ. After detrending both time series by first differencing, the contemporary correlations between realized profitability and realized spreads for fast stocks lie at below 0.3 regardless of the horizon chosen for the estimation. For slow stocks, the correlations are much higher, hovering around 0.5. Still, these numbers suggest there is significant variation in realized profitability in the time series not captured by the conventional realized spread measure, even for the slow stocks.

2.6 Robustness: Alternative Inventory Tracking

When using FIFO, the aggregate term structure for the realized profitability is downward sloping. The issue is that, as we previously discussed, the downward slope may be a mechanical artifact of the interaction of the FIFO system with order imbalance. In Figure 24 we plot out the empirical term structure under FIFO for different stock-day trade imbalance deciles. Consistent with our result from Section B we observe a downward-sloping term structure that gets more dramatic as the level of order imbalance increases. Interestingly is that when restricting ourselves to low-imbalance stock days, when the influence of the mechanical bias is lower, the term structure is upward-sloping, consistent with our results using LIFO.

In Figure 25 we perform the slow-fast τ sorts using all stock days (top) and the 25% stock days with the lowest imbalance (bottom). The main difference in the term structure under FIFO when including high imbalance days seems to be one of level as they are all downward sloping. Restricting ourselves to low imbalance days, we get patterns broadly consistent with the LIFO results.

The shape of the LIFO term structure is stable across stock days with low or high order imbalances whereas the FIFO term structure is not. At first glance, this raises the concern that FIFO is capturing something LIFO is not on high-imbalance stock days. This behavior in the FIFO term structure is however perfectly in line with the mechanical relationship between the FIFO term structure and order imbalance examined in Section B. In other words, we believe that the drastic change in the FIFO term structure is due to a statistical artifact inherent to the method itself. Any alternative explanation would have to argue for the existence of an economically significant factor affecting liquidity provider inventories on high imbalance stock days that: (1) reverses the risk-return trade-off observed in low imbalance days, (2) is distinct from

the FIFO mechanical bias, and (3) is sensitive to measurement methodology, showing up in FIFO but not LIFO.

2.7 Conclusion

The conventional realized spread estimates a mark-to-market profit at a prespecified (exogenous) market-making horizon; this profit can deviate significantly from the profits to liquidity provision if the price subsequently moves at the time of the exit. By tracking the cumulative inventory positions of all passive liquidity providers in the US equity market and matching each position with its offsetting trade, we construct a measure of profits to liquidity provision (realized profitability) that matches the dynamics of Virtu's market-making revenue much better than realized spread (at any reasonably prespecified horizon).

To make sense of the difference between our realized profitability and conventional realized spread, we assess how realized profitability varies with the endogenous market-making horizon τ and compare that to the term structure of realized spread in Conrad and Wahal (2020). We find, unlike the conventional realized spread, which is sharply decreasing in τ, our realized profitability is strictly increasing in τ. Since longer inventory turnaround typically implies a higher risk of market making, we interpret our result as consistent with the risk-return trade-off faced by an average liquidity provider in the competitive market-making business. By decomposing our realized profitability into an alternative realized spread component (measured with endogenized τ for each trade) and the effective spread at the exit trade, we show the bias in the conventional realized spread as a proxy for market-making profit is mainly caused by the specification of common τ across all trades.

Figure 1: ISO Single Limit-Order Book (SLOB) Execution

The top two panels reflect the Bid side of the order books on two exchanges, A (blue) and B (orange), with protected bids in bold. An ISO implemented trade looking to climb up the exchange A book beyond the first level would have to commit to clearing out the exchange B protected quote. The ISO exemption only allows for orders to trade-through better-priced protected quotes if the trader commits to concurrently clearing out those quotes. An ISO implemented trade looking to climb up the exchange A book would execute as-if there was a single trade venue with the liquidity collected in the bottom-left order book. This is in contrast to the bottom-right book constructed out of all the available liquidity from both exchanges; note that the ISO-SLOB misses the liquidity available at \$4.5. After the protected quotes are cleared out, the NBB would update to the quote for 50 @ \$4.5 on exchange B.

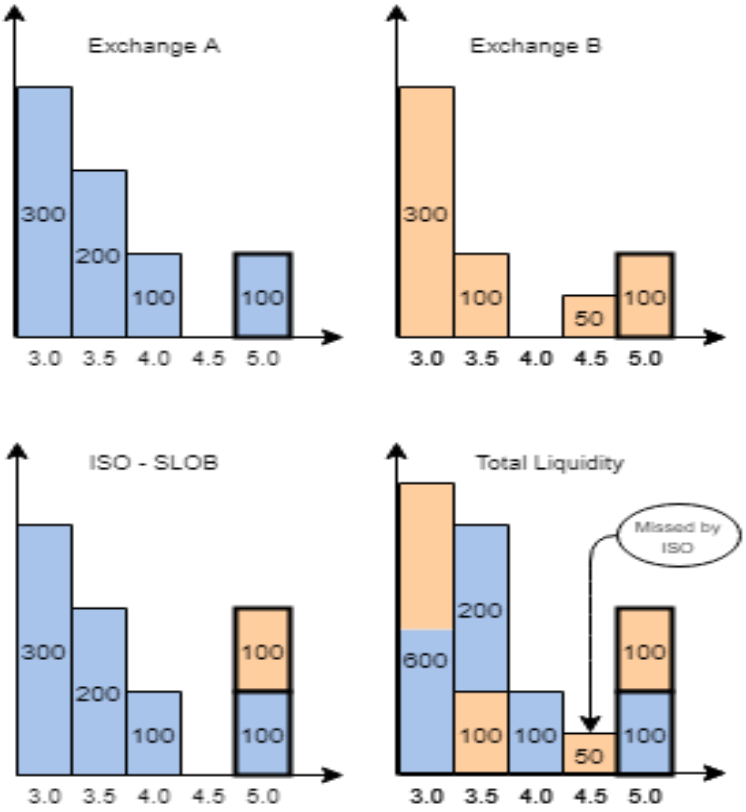


Table 1: Summary Statistics

Summary statistics for a variety of sub-samples are reported below. The sample is comprised of 2,657 publicly traded securities for the period spanning Jan 2019 - Apr 2021 and covers 93% of the total traded dollar volume. The top two panels report values for small, medium, and large sized stocks along with values for foreign securities, REITs, and ETFs. The bottom two panels break out the sample into sub-samples by the median number of public trade venues for each security.

	Small (20%)	Mid (60%)	Big (20%)	ALL
Number of Securities	391	1170	390	2651
Market Cap (Billions $)	0.25	2.93	63.18	14.05
Proportion Sweep Volume	43.84%	40.02%	44.54%	47.78%
Median Number of Exchanges	2	5	8	6
Effective Spread (bps)	15.55	5.46	1.90	3.61
Relative ISO cps (bps)	5.30	2.27	0.76	1.36
RISO cps / Effective Spread	0.34	0.42	0.41	0.41

	Common Stock	Foreign	REIT	ETF
Number of Securities	1951	190	89	354
Market Cap (Billions $)	14.43	17.07	11.77	11.34
Proportion Sweep Volume	43.71%	44.63%	42.01%	57.17%
Median Number of Exchanges	5	8	7	5
Effective Spread (bps)	3.82	4.95	3.65	1.36
Relative ISO cps (bps)	1.53	2.11	1.55	0.45
RISO cps / Effective Spread	0.40	0.43	0.44	0.33

	≤ 3 venues	4	5	6
Number of Securities	584	1170	390	1951
Market Cap (Billions $)	1.57	3.34	4.38	6.84
Proportion Sweep Volume	52.10%	47.95%	43.88%	44.04%
Effective Spread (bps)	8.97	4.93	5.07	4.02
Relative ISO cps (bps)	2.52	1.81	1.91	1.56
RISO cps / Effective Spread	0.28	0.37	0.38	0.39

	7	8	> 9 venues	ALL
Number of Securities	190	89	354	2651
Market Cap (Billions $)	13.15	22.32	60.12	14.05
Proportion Sweep Volume	46.90%	46.37%	47.33%	47.78%
Effective Spread (bps)	3.04	2.45	3.56	3.61
Relative ISO cps (bps)	1.23	1.00	1.52	1.36
RISO cps / Effective Spread	0.40	0.41	0.43	0.41

Figure 2: Non-ISO and ISO Trade Mechanics

MO trade implementations execute in a sequential manner; orders can only execute at venues with the best prices and must wait for NBBO quotes to update before the next leg of the trade can execute. When using a sequence of MOs, the trader must (1) route the MO to an exchange quoting at the NBBO price. After receiving a MO (2) the exchange checks the tape to see if they are able to fill the order, if it does, it then updates the tape. In order to ensure that subsequent MOs are routed correctly, the trader waits to observe the updated quotes (3) before repeating the process until the trade is complete. ISO trade implementations can trade-through the NBBO and execute across multiple venues simultaneously. ISOs are routed across multiple venues (1) at the same time, exchanges fill the ISOs and update the tape (2).

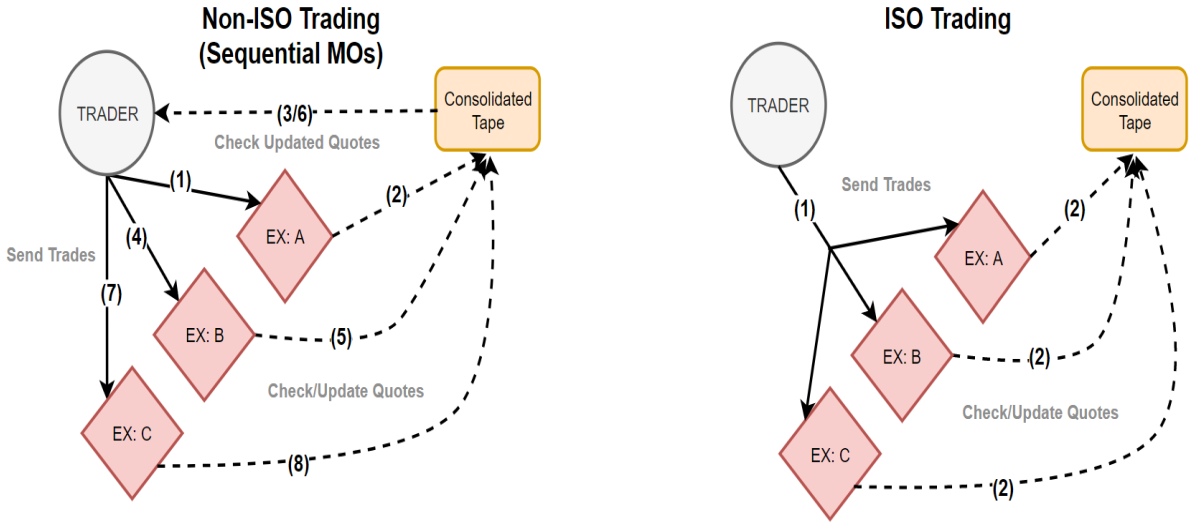


Figure 3: Trading Example

ISO and MO implementations of a sale for 200 shares, starting at time t are compared in an environment with three exchanges (A, B, and C). Posted quotes are expressed as (# Shares available @ Price-per-share). Protected quotes are colored with a blue fill, unprotected quotes are colored with a green fill, and hidden orders are denoted with a light gray font). The ISO implementation consists of 3 sales simultaneously executed across the three exchanges at time t. The first MO implementation assumes no change in the posted quotes as the three orders at times t, $t+1$, and $t+2$ execute. MOs can only execute at the best available price, after the liquidity at \$5 is cleared out on exchanges A and B, the NBB updates to the quote at \$4.5 on exchange A. The second MO implementation assumes that the bid quotes are revised down to the prices in red after the first MO execution of 100 @ 5 is observed.

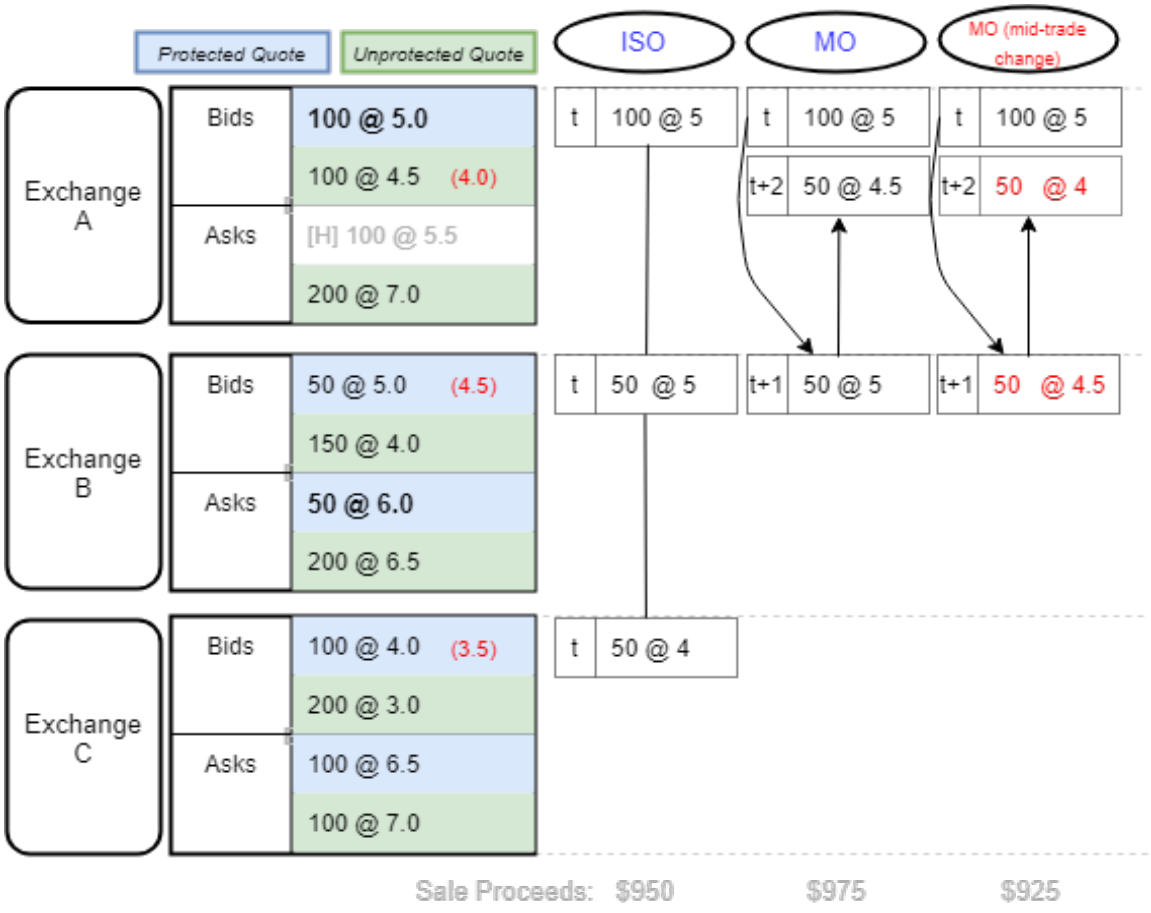


Figure 4: ISO and MO Order-Book Executions

Two books are constructed using the posted quotes from the three exchanges from Figure 3. The pseudo-book is constructed only from the prevailing protected quotes (blue) across the exchanges; these quotes are level 1 (L1) quotes corresponding to the top of each exchange's book. A ISO trades against this pseudo book the same as a market order would, instantaneously climbing the order-book. The combined book is a hypothetical order book constructed from both protected and unprotected (green) quotes from the three exchanges. An MO implementation would execute similar to 3 sequential market orders trading against the combined book, though with pauses between MOs to allow the NBBO to update. During these pauses the posted liquidity could adversely move against the investor mid-trade.

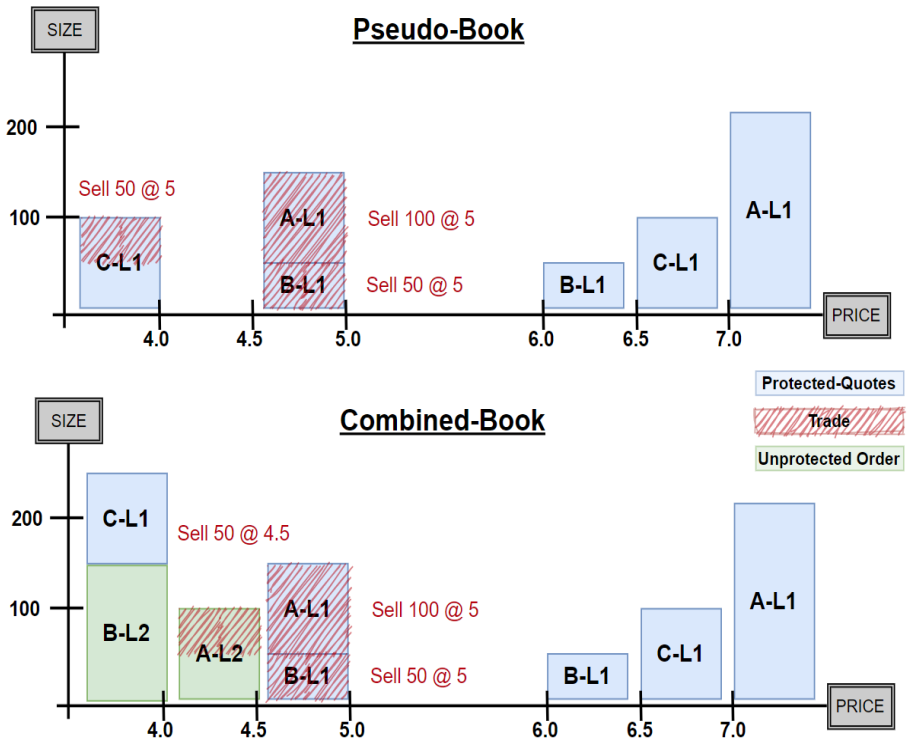


Figure 5: Example Trade And Quote Ordering 1

Below is a possible presentation of the merged quote and trade updates for the ISO implemented sale of 200 shares from section 1.2. Quotes (blue) on the left and order executions (red) on the right are sorted by the order reported on the consolidated tape. The first three quote updates establish the prevailing protected bids, with a NBB of \$5, immediately preceding the initiation of the ISO sale. The first ISO clears out exchange A's protected bid of 100 @ 5. This leads to the updated protected quote of 100 @ 4.5 for exchange A, this quote becomes the NBB after exchange B's protected quote is cleared out by the second ISO. The first two ISOs have TECs of zero because the sales occurred at the NBB. By the time the last ISO on exchange C is reported, the prevailing NBB is 4.5, leading to a TEC of 50 $shares \times (4.5 - 4) = \$25$.

	Quote Updates				Order Executions				
Order	Exchange	Bid Size	Bid	NBB	Exchange	Type	Order Size	Price	TEC
1	A	100	5	5					
2	B	50	5	5					
3	C	100	4	5					
4					A	ISO	100	5	0
5	A	100	4.5	5					
6					B	ISO	50	5	0
7	B	150	4	4.5					
8					C	ISO	50	4	25

Figure 6: Example Trade And Quote Ordering 2

Below is another possible presentation of the merged quote and trade updates for the ISO implemented sale of 200 shares from section 1.2. Quotes (blue) on the left and order executions (red) on the right are sorted by the order reported on the consolidated tape. The first three quote updates establish the prevailing protected bids, with a NBB of \$5, immediately preceding the initiation of the ISO sale. In this scenario, the executions of all three ISO legs of the sale are reported before the NBB is updated. As in the ordering in Figure 5, the ISOs on exchange A and B have a TEC of zero as the sales occured at the NBB. The difference here is that because the prevailing NBB was \$5 (instead of \$4.5) when the exchange C ISO was reported, that order has a TEC of 50 *shares* $\times(5-4) = \$50$, which is an overestimate of the actual excess costs of \$25.

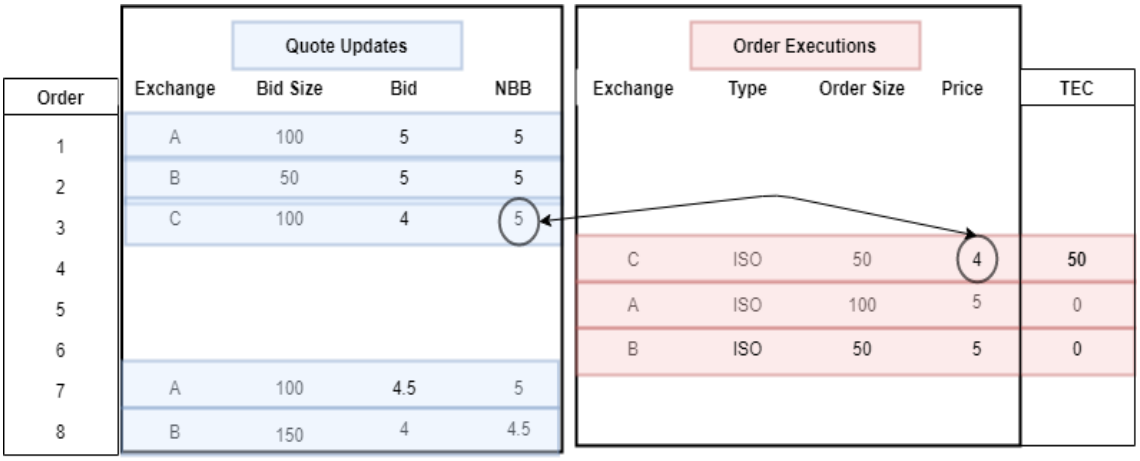


Figure 7: Relative ISO Costs-Per-Share Over Time

The trade dollar volume weighted average of the relative ISO costs per share to average effective spread ratio using the whole sample is plotted across time in blue. The red line is simply the time-series average of the blue.

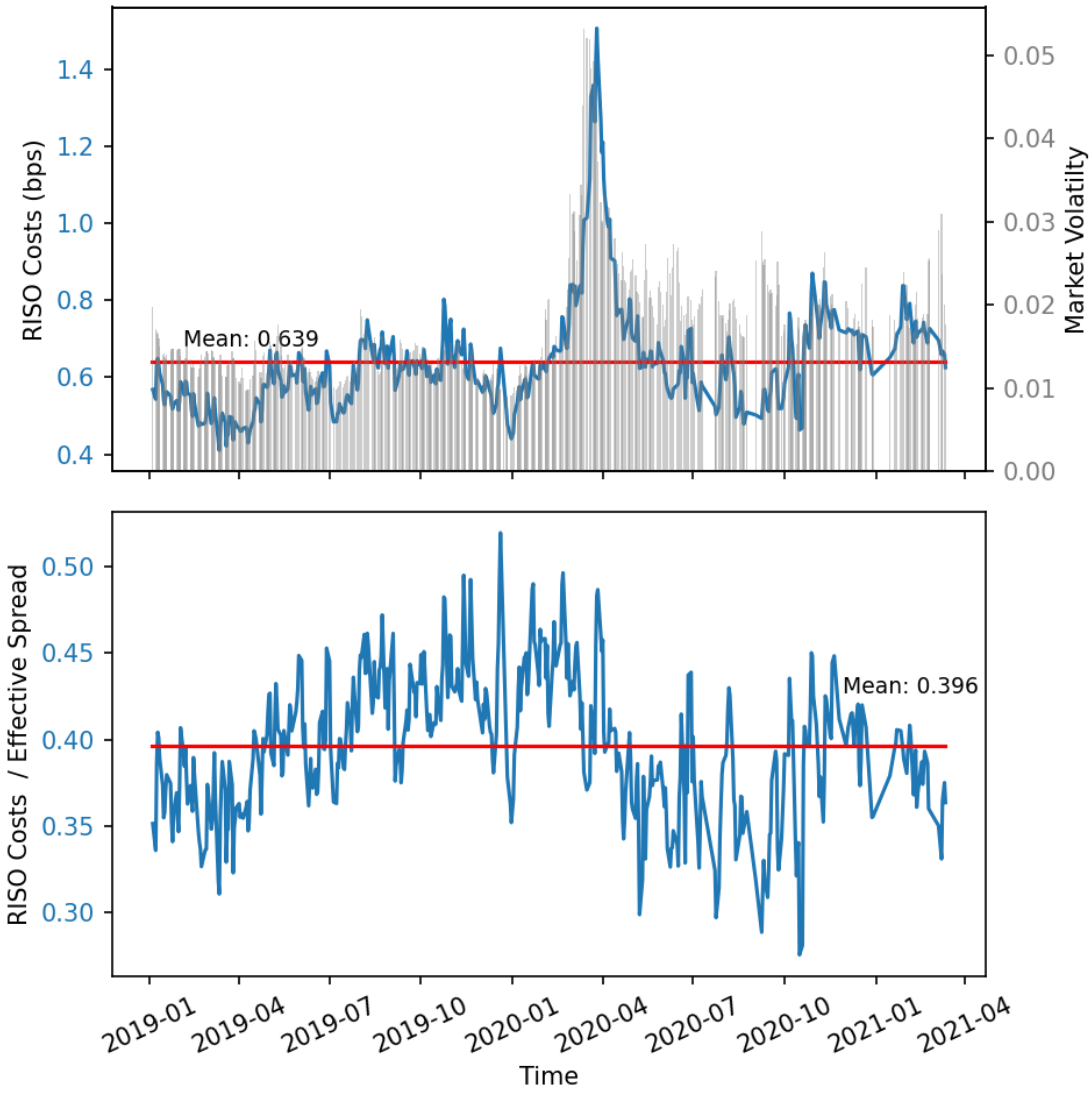


Figure 8: Neighbor Trade Comparison vs. Daily Aggregate

Below is a graphical comparison of the daily aggregation and neighbor trade methods of estimating the ISO premium. Consecutive individual orders are grouped into either a NISO (blue) or ISO (red) trade block. Trade blocks are chronologically ordered (left to right) for the trade day from the open to close. The top panel illustrates the daily aggregation method. All ISO trade blocks across the day are collected together to calculate the ISO costs per share (ISOcps), similarly all NISO trade blocks throughout the day are collected to calculate the NISO costs per share. The difference, ISOcps - NISOcps = RISOcps is the day's relative ISO costs per share and comprises the daily aggregate measure o the ISO premium. The second half illustrates the neighboring trade method. Each ISO trade block with an immediately preceding NISO block is paired up with that NISO block to compute $NISOcps_i$, the difference in the costs per-share between the two blocks. The volume weighted $NISOCPS_i$ measures are summed up to get, $NISOcps$, the neighboring trades derived measure of the ISO premium.

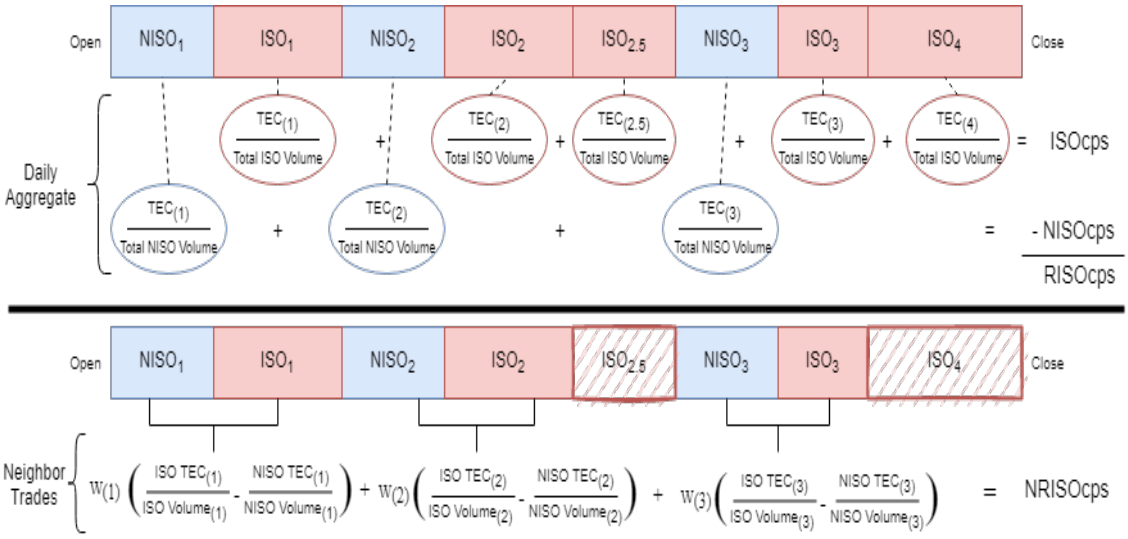


Table 2: Previous Fragmentation Measures

Regression results for the log-log panel regressions of the form:

$$\ln y_{i,t} = \alpha + \beta' \ln X_{i,t} + \epsilon_{i,t}$$

for three subsamples are reported. The response variable y is either a realized spread, effective spread, ISO TEC costs-per-share, or relative ISO TEC costs-per-share; all measured in basis-points. Here the design vector X is comprised of $(1 - HHI)$ and $\frac{DarkVol}{TotalVol}$. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Month-time fixed effects are included in all specifications.

Panel A: All Securities										
Variable	Realized Spreads		Effective Spreads		Gross ISO Costs		Relative ISOcps		Scaled RISOcps	
$Constant$	-0.94^{***}	0.07^{***}	0.55^{***}	0.09^{***}	0.73^{***}	1.20^{***}	-0.12^{***}	0.78^{***}	-0.69^{***}	-0.51^{***}
	(0.05)	(0.01)	(0.06)	(0.01)	(0.06)	(0.04)	(0.06)	(0.02)	(0.02)	(0.01)
$(1 - HHI)$	-3.50^{***}	0.05	-2.62^{***}	0.14^{***}	-1.20^{***}	0.65^{***}	-2.59^{***}	0.85^{***}	-0.19^{***}	0.82^{***}
	(0.12)	(0.04)	(0.13)	(0.04)	(0.14)	(0.04)	(0.14)	(0.04)	(0.04)	(0.03)
$\frac{DarkVol}{TotalVol}$	-0.09^{***}	1.52^{***}	-0.01	0.14^{***}	0.16^{***}	0.15^{***}	0.34^{***}	0.37^{***}	0.32^{***}	0.26^{***}
	(0.03)	(0.01)	(0.04)	(0.01)	(0.04)	(0.08)	(0.04)	(0.01)	(0.01)	(0.01)
Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Panel B: Median Public Venues ≤ 3										
Variable	Realized Spreads		Effective Spreads		Gross ISO Costs		Relative ISOcps		Scaled RISOcps	
$Constant$	-0.30^{***}	0.92^{***}	0.65^{***}	2.08^{***}	0.42^{***}	1.66^{***}	-0.02^{***}	1.43^{***}	-0.56^{***}	-0.53
	(0.10)	(0.02)	(0.12)	(0.01)	(0.12)	(0.02)	(0.12)	(0.02)	(0.04)	(0.02)
$(1 - HHI)$	-1.85^{***}	0.03	-1.56^{***}	0.01	-1.17^{***}	0.34^{***}	-1.48^{***}	0.38^{***}	0.05	0.39^{***}
	(0.14)	(0.05)	(0.20)	(0.03)	(0.16)	(0.04)	(0.17)	(0.05)	(0.05)	(0.04)
$\frac{DarkVol}{TotalVol}$	-0.48^{***}	0.07^{***}	-0.75^{***}	0.07^{***}	-0.56^{***}	0.10^{***}	-0.52^{***}	0.26^{***}	0.26^{***}	0.12^{***}
	(0.06)	(0.01)	(0.07)	(0.01)	(0.07)	(0.01)	(0.07)	(0.01)	(0.02)	(0.01)
Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Panel B: Median Public Venues > 9										
Variable	Realized Spreads		Effective Spreads		Gross ISO Costs		Relative ISOcps		Scaled RISOcps	
$Constant$	-0.12	-0.43^{***}	2.09^{***}	1.21^{***}	2.42^{***}	1.41^{***}	1.35^{***}	0.72^{***}	-0.75^{***}	-0.48^{***}
	(0.22)	(0.07)	(0.25)	(0.06)	(0.26)	(0.07)	(0.25)	(0.02)	(0.08)	(0.05)
$(1 - HHI)$	-1.46^{***}	0.90^{***}	2.41^{***}	0.76^{***}	4.10^{***}	1.68^{***}	3.19^{***}	2.32^{***}	0.63^{***}	1.54^{***}
	(0.56)	(0.27)	(0.67)	(0.20)	(0.70)	(0.24)	(0.71)	(0.24)	(0.23)	(0.15)
$\frac{DarkVol}{TotalVol}$	0.94^{***}	0.53^{***}	0.73^{***}	0.25^{***}	0.76^{***}	0.31^{***}	0.41^{***}	0.50^{***}	0.18^{***}	0.25^{***}
	(0.14)	(0.05)	(0.16)	(0.03)	(0.16)	(0.04)	(0.15)	(0.04)	(0.05)	(0.03)
Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Figure 9: RV Dispersion Illustration

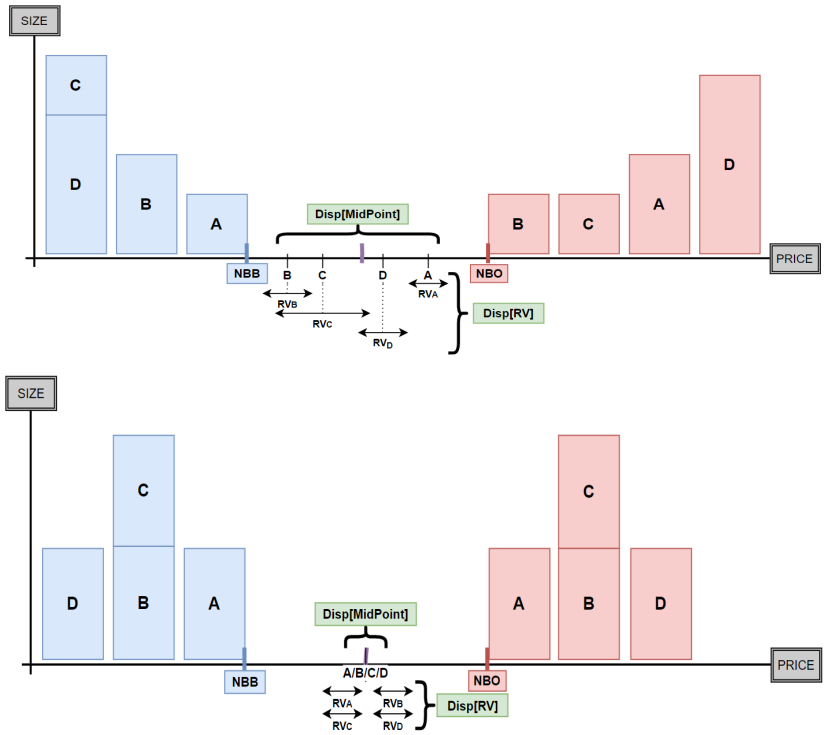


Figure 10: Trade and Midpoint RV Estimates

The log trade dollar volume weighted average RV estimates across time for the subset of 50 heavily traded securities are plotted across time. The blue line is derived using RV estimates calculated using transaction values and the orange line is derived using RV estimates calculated using midpoint shifts. Appendix A describes the RV estimation procedure in depth.

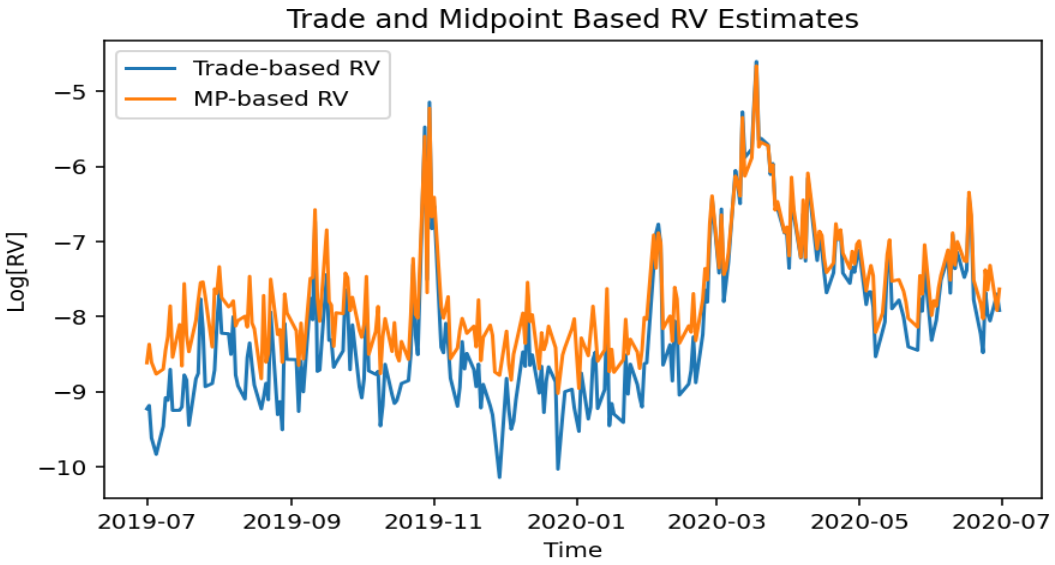


Table 3: Dispersion-Based Fragmentation Measures

Regression results for the log-log panel regressions of the form:

$$\ln y_{i,t} = \alpha + \beta' \ln X_{i,t} + \epsilon_{i,t}$$

for three subsamples are reported. The response variable y is either a realized spread, effective spread, ISO TEC costs-per-share, or relative ISO TEC costs-per-share; all measured in basis-points. Here the design vector X is comprised of $Disp[RV]$, AOS, and RV^{NBBO}. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Month-time fixed effects are included in all specifications.

Panel A: All Securities										
Variable	Realized Spreads		Effective Spreads		Gross ISO Costs		Relative ISOcps		Scaled RISOcps	
Constant	1.07^{***}	1.67^{***}	3.72^{***}	3.27^{***}	4.65^{***}	3.31^{***}	2.04^{***}	2.02^{***}	-1.53^{***}	-1.15^{***}
	(0.10)	(0.07)	(0.09)	(0.05)	(0.08)	(0.06)	(0.10)	(0.06)	(0.05)	(0.04)
$Disp[RV]$	0.25^{***}	0.10^{***}	0.31^{***}	0.12^{***}	0.35^{***}	0.13^{***}	0.33^{***}	0.13^{***}	0.02^{***}	0.02^{***}
	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
AOS	0.36^{***}	0.09^{***}	0.30^{***}	0.03^{***}	0.08^{***}	0.01	0.34^{***}	0.05^{***}	0.07^{***}	0.03^{***}
	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
RV^{NBBO}	-0.11^{***}	0.08^{***}	-0.08^{***}	0.06^{***}	-0.10^{***}	0.07^{***}	-0.17^{***}	0.03^{***}	-0.06^{***}	-0.03^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Panel B: Median Public Venues ≤ 3										
Variable	Realized Spreads		Effective Spreads		Gross ISO Costs		Relative ISOcps		Scaled RISOcps	
Constant	1.73^{***}	1.59^{***}	2.99^{***}	3.11^{***}	3.07^{***}	2.88^{***}	1.28^{***}	1.65^{***}	-1.60^{***}	-1.32^{***}
	(0.18)	(0.12)	(0.15)	(0.06)	(0.15)	(0.07)	(0.17)	(0.09)	(0.09)	(0.07)
$Disp[RV]$	0.08^{***}	0.03^{***}	0.11^{***}	0.03^{***}	0.20^{***}	0.05^{***}	0.17^{***}	0.05^{***}	0.05^{***}	0.02^{***}
	(0.02)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.00)
AOS	0.26^{***}	0.12^{***}	0.44^{***}	0.09^{***}	0.30^{***}	0.05^{***}	0.42^{***}	0.06^{***}	0.04^{***}	0.00
	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)
RV^{NBBO}	0.12^{***}	0.11^{***}	0.17^{***}	0.14^{***}	0.04^{***}	0.13^{***}	-0.02^{***}	0.04^{***}	-0.14^{***}	-0.08^{***}
	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Panel B: Median Public Venues > 9										
Variable	Realized Spreads		Effective Spreads		Gross ISO Costs		Relative ISOcps		Scaled RISOcps	
Constant	2.92^{***}	2.74^{***}	5.44^{***}	4.21^{***}	6.15^{***}	4.63^{***}	4.02^{***}	3.22^{***}	-1.17^{***}	-0.82^{***}
	(0.22)	(0.16)	(0.17)	(0.15)	(0.15)	(0.16)	(0.20)	(0.18)	(0.12)	(0.11)
$Disp[RV]$	0.19^{***}	0.21^{***}	0.41^{***}	0.28^{***}	0.47^{***}	0.32^{***}	0.44^{***}	0.31^{***}	0.03^{**}	0.05^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
AOS	0.10^{***}	0.02^{***}	0.02	0.02^{**}	-0.08^{***}	-0.04	0.07^{***}	0.04^{***}	0.03	0.06^{***}
	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
RV^{NBBO}	0.14^{***}	0.06^{***}	-0.13^{***}	0.08^{***}	-0.18^{***}	-0.09^{***}	-0.21^{***}	0.10^{***}	-0.06^{**}	-0.03^{**}
	(0.05)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)	(0.02)	(0.02)	(0.01)
Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Figure 11: Log and Level Regression Residuals

Standardized histograms (with skewness) for the residuals for the panel regressions of the form:

$$RISOcps_{i,t} = \alpha + \beta_1' \ln F_{i,t} + \beta_2' \ln X_{i,t} + \epsilon_{i,t}$$

Where the response variable is either log transformed or not. Fragmentation measures are comprised of $F_{i,t} = \left[\text{Disp}[RV]_{i,t}, AOS_{i,t}, (1 - HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}\right]$. Control variables captured in the design vector X are $\{RV^{nbbo}, Spread, sweepProp, ILLIQ\}$. Month-time and security fixed effects are included in both specifications. The level regression residual histogram is shaded in red and the one for the log regression is in blue, the black dashed-line denotes a benchmark standard-normal probability distribution function.

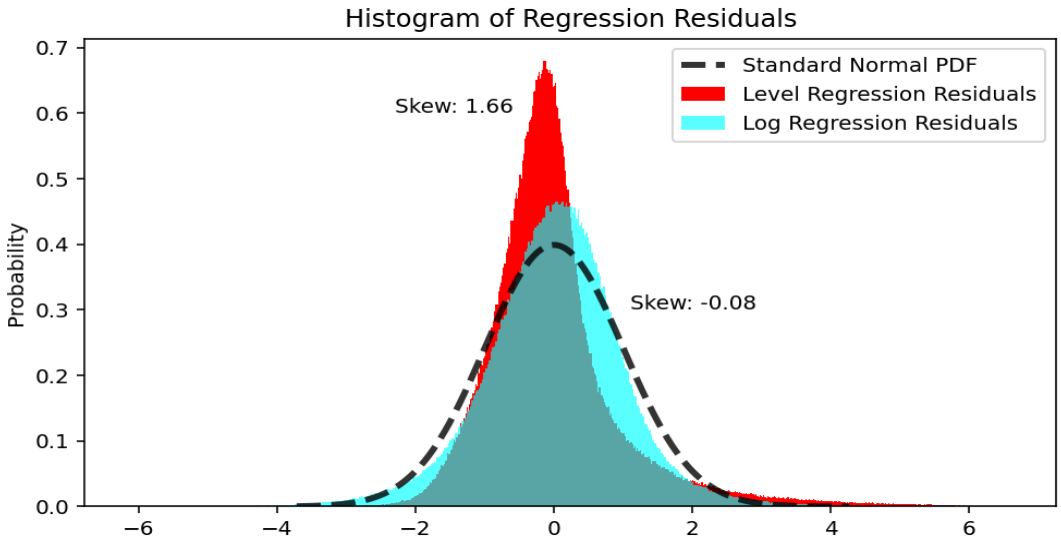


Table 4: Results Across Subsamples

Regression results for the log-log panel regressions of the form:

$$\ln RISOcps_{i,t} = \alpha + \beta_1' \ln F_{i,t} + \beta_2' \ln X_{i,t} + \epsilon_{i,t}$$

for all securities, the top 50 securities, and those securities which are traded on 7 or more exchanges. Fragmentation measures are comprised of $F_{i,t} = \left[\text{Disp}[RV]_{i,t}, AOS_{i,t}, (1 - HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}\right]$. Control variables captured in the design vector X are $\{RV^{nbbo}, Spread, sweepProp, ILLIQ\}$. The first column for each sample reports the regression results excluding the control variable X from the specification. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Month-time and security fixed effects are included in all specifications.

Variable	Relative ISO Costs Per Share					
	All		≤ 3 Trade Venues		> 9 Trade Venues	
Constant	2.637***	5.009***	2.069***	4.224***	3.801***	6.859***
	(0.067)	*(0.008)*	*(0.087)*	*(0.118)*	*(0.203)*	*(0.239)*
Disp[*RV*]	0.143***	0.176***	0.069***	0.073***	0.239***	0.377***
	(0.004)	*(0.006)*	*(0.005)*	*(0.007)*	*(0.013)*	*(0.017)*
AOS	0.030***	−0.014***	0.047***	−0.020**	0.026***	0.003
	(0.004)	*(0.003)*	*(0.009)*	*(0.008)*	*(0.011)*	*(0.009)*
$(1 - HHI)$	1.157***	1.277***	0.623***	0.679***	2.252***	2.303***
	(0.040)	*(0.036)*	*(0.062)*	*(0.063)*	*(0.189)*	*(0.161)*
$\frac{DarkVol}{TotalVol}$	0.291***	0.271***	0.226***	0.198***	0.240***	0.296***
	(0.008)	*(0.007)*	*(0.013)*	*(0.014)*	*(0.033)*	*(0.029)*
RV^{NBBO}		−0.064***		−0.074***		−0.186***
		(0.006)		*(0.010)*		*(0.017)*
QuotedSpread		0.545***		0.711***		0.478***
		(0.010)		*(0.029)*		*(0.025)*
SweepProportion		0.065***		0.084***		0.116***
		(0.008)		*(0.013)*		*(0.030)*
ILLIQ		0.017***		0.020***		0.018***
		(0.002)		*(0.004)*		*(0.003)*
Entity-Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month-Effects	Yes	Yes	Yes	Yes	Yes	Yes
R2	21.10	30.03	11.19	11.95	45.35	50.32
Num. Obs	1,10,530		132,266		114,100	
Num. Entities	2,637		584		238	

Table 5: Results Across Small/Large Subsamples

Regression results for the log-log panel regressions of the form:

$$\ln RISOcps_{i,t} = \alpha + \beta_1' \ln F_{i,t} + \beta_2' \ln X_{i,t} + \epsilon_{i,t}$$

for all securities and the market-cap size quintile subsets. Fragmentation measures are comprised of $F_{i,t} = \left[\text{Disp}[RV]_{i,t}, AOS_{i,t}, (1-HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}\right]$. Control variables captured in the design vector X are $\{RV^{nbbo},\ Spread,\ sweepProp,\ ILLIQ\}$. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Month-time and security fixed effects are included in all specifications.

Variable	Relative ISO Costs Per Share					
	Smallest 20%		Middle 60%		Largest 20%	
Constant	2.936^{***}	5.964^{***}	2.944^{***}	5.377^{***}	2.763^{***}	4.288^{***}
	(0.164)	*(0.178)*	*(0.082)*	*(0.096)*	*(0.161)*	*(0.157)*
Disp[*RV*]	0.131^{***}	0.236^{***}	0.146^{***}	0.173^{***}	0.179^{***}	0.148^{***}
	(0.010)	*(0.015)*	*(0.005)*	*(0.007)*	*(0.010)*	*(0.011)*
AOS	0.094^{***}	0.019	0.027^{***}	-0.041^{***}	0.059^{***}	0.004
	(0.019)	*(0.015)*	*(0.007)*	*(0.006)*	*(0.009)*	*(0.008)*
$(1 - HHI)$	0.738^{***}	0.781^{***}	1.257^{***}	1.384^{***}	1.471^{***}	1.648^{***}
	(0.096)	*(0.089)*	*(0.049)*	*(0.045)*	*(0.098)*	*(0.101)*
$\frac{DarkVol}{TotalVol}$	0.193^{***}	0.335^{***}	0.247^{***}	0.256^{***}	0.339^{***}	0.308^{***}
	(0.020)	*(0.022)*	*(0.008)*	*(0.009)*	*(0.022)*	*(0.019)*
RV^{NBBO}		-0.194^{***}		-0.049^{***}		-0.016
		(0.014)		*(0.008)*		*(0.011)*
QuotedSpread		0.416^{***}		0.507^{***}		0.546^{***}
		(0.024)		*(0.012)*		*(0.022)*
SweepProportion		0.212^{***}		0.093^{***}		0.096^{***}
		(0.021)		*(0.013)*		*(0.023)*
ILLIQ		0.055^{***}		0.021^{***}		-0.007^{**}
		(0.005)		*(0.003)*		*(0.003)*
Entity-Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month-Effects	Yes	Yes	Yes	Yes	Yes	Yes
R2	9.52	15.08	8.05	19.16	11.76	23.63
Num. Obs	96,963		521,735		185,553	
Num. Entities	388		1,163		388	

Table 6: Regression Results With and Without COVID-19

Regression results for the log-log panel regressions of the form:

$$\ln RISOcps_{i,t} = \alpha + \beta_1' \ln F_{i,t} + \beta_2' \ln X_{i,t} + \epsilon_{i,t}$$

for all securities for the whole sample period, the pre-COVID19, and COVID19-forward sample periods. Pre-COVID19 period spans from Jan 1, 2019 to Feb 1, 2020; the COVID19-Forward runs from Feb 1, 2020 to Apr 30, 2021. Fragmentation measures are comprised of $F_{i,t} = \left[\text{Disp}[RV]_{i,t}, AOS_{i,t}, (1 - HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}\right]$. Control variables captured in the design vector X are $\{RV^{nbbo}, Spread, sweepProp, ILLIQ\}$. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Month-time and security fixed effects are included in all specifications.

Variable	Relative ISO Costs Per Share					
	All		Pre-COVID19 Period		COVID19 Plus Period	
Constant	2.637***	5.009***	1.993***	4.221***	2.799***	4.664***
	(0.067)	*(0.008)*	*(0.065)*	*(0.092)*	*(0.098)*	*(0.107)*
Disp[*RV*]	0.143***	0.176***	0.110***	0.091***	0.147***	0.155***
	(0.004)	*(0.006)*	*(0.004)*	*(0.005)*	*(0.006)*	*(0.007)*
AOS	0.030***	−0.014***	0.026***	−0.011**	0.041***	−0.011**
	(0.004)	*(0.003)*	*(0.004)*	*(0.004)*	*(0.006)*	*(0.005)*
$(1 - HHI)$	1.157***	1.277***	1.089***	1.201***	1.186***	1.237***
	(0.040)	*(0.036)*	*(0.043)*	*(0.042)*	*(0.051)*	*(0.044)*
$\frac{DarkVol}{TotalVol}$	0.291***	0.271***	0.252***	0.239***	0.309***	0.269***
	(0.008)	*(0.007)*	*(0.008)*	*(0.008)*	*(0.013)*	*(0.012)*
RV^{NBBO}		−0.064***		−0.006		−0.076***
		(0.006)		*(0.005)*		*(0.008)*
QuotedSpread		0.545***		0.630***		0.602***
		(0.010)		*(0.014)*		*(0.014)*
SweepProportion		0.065***		0.119***		0.066***
		(0.008)		*(0.009)*		*(0.011)*
ILLIQ		0.017***		0.001		0.009***
		(0.002)		*(0.003)*		*(0.003)*
Entity-Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month-Effects	Yes	Yes	Yes	Yes	Yes	Yes
R2	21.10	30.03	13.65	17.70	19.65	22.90
Num. Obs	1,10,530		520,085		442,053	
Num. Entities	2,637		2,637		2,637	

Table 7: Neighbor-Trade Comparison Results

Regression results for the log-log panel regressions of the form:

$$\ln y_{i,t} = \alpha + \beta_1' \ln F_{i,t} + \beta_2' \ln X_{i,t} + \epsilon_{i,t}$$

for all securities, where the response variable is a measure of the relative ISO costs per share calculated using daily means or by comparing neighboring trades . Fragmentation measures are comprised of $F_{i,t} = \Big[\text{Disp}[RV]_{i,t}, MPD_{i,t}, AOS_{i,t}, (1 - HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}\Big]$. Control variables captured in the design vector X are $\{RV^{nbbo}, Spread, sweepProp, ILLIQ\}$. The first column of each method reports the regression results excluding the control variable X from the specification. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Month-time and security fixed effects are included in all specifications.

Variable	*Relative ISO Costs Per Share*			
	Daily Aggregate		Neighboring Trades	
Constant	2.721***	5.105***	3.663***	5.789***
	(0.063)	*(0.077)*	*(0.060)*	*(0.078)*
Disp[*RV*]	0.148***	0.183***	0.170***	0.143***
	(0.004)	*(0.005)*	*(0.004)*	*(0.005)*
AOS	0.013*	−0.065***	−0.003	−0.050***
	(0.008)	*(0.006)*	*(0.006)*	*(0.006)*
$(1 - HHI)$	1.071***	1.218***	0.474***	0.698***
	(0.040)	*(0.035)*	*(0.040)*	*(0.037)*
$\frac{DarkVol}{TotalVol}$	0.291***	0.275***	0.210***	0.317***
	(0.008)	*(0.007)*	*(0.007)*	*(0.008)*
RV^{NBBO}		−0.075***		0.022***
		(0.006)		*(0.005)*
QuotedSpread		0.541***		0.378***
		(0.009)		*(0.011)*
SweepProportion		0.066***		0.352***
		(0.008)		*(0.009)*
ILLIQ		0.019***		0.009***
		(0.002)		*(0.002)*
Entity-Effects	Yes	Yes	Yes	Yes
Month-Effects	Yes	Yes	Yes	Yes
R2	21.99	30.25	37.52	30.24
Num. Obs	1,069,019		873,098	

Table 8: 2-Stage Heckman Correction

Regression results for the 2-stage Heckman procedure for correcting selection bias are reported below. The first stage-probit specification follows:

$$\Phi^{-1}(ISOpercent_{i,t}) = Z_{i,t}\gamma + u_{i,t}$$

and the second-stage specification:

$$ISOcosts_{i,t} = X_{i,t}\beta + \hat{\lambda}_{i,t}\theta + \epsilon_{i,t}$$

where $\hat{\lambda}_{i,t} = \frac{\phi(Z_{i,t}\hat{\gamma})}{\Phi(Z_{i,t}\hat{\gamma})}$ denotes the estimated inverse Mills ratio from the first stage. Design variables included in the first and second stages are $\Big[\text{Disp}[RV]_{i,t-1}, RV^{nbbo}_{i,t-1}, Spread_{i,t-1}, (1-HHI_{i,t-1}),\Big]$ and $\Big[\text{Disp}[RV]_{i,t}, (1-HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}, RV^{nbbo}_{i,t}, Spread_{i,t}, \hat{\lambda}_{i,t}\Big]$ respectively. In this specify time periods are split into consecutive 15-min trading periods. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Day-time and security fixed effects are included in all specifications.

Stage 1 Probit		Stage 2 OLS		
Variables	$\Phi^{-1}(ISOpercent)$	Variables	Relative ISO Costs	
$Disp[RV]_{t-1}$	-0.003***	$Constant$	-1.848***	-1.924***
	(.001)		*(0.025)*	*(0.021)*
RV^{NBBO}_{t-1}	0.013***	$Disp[RV]_t$	0.010***	0.008***
	(.001)		*(0.002)*	*(0.001)*
$(1-HHI_{t-1})$	-0.142***	$(1-HHI_t)$	0.656***	0.579***
	(.008)		*(0.018)*	*(0.016)*
QuoteSpread_{t-1}	-0.100***	$\frac{DarkVol}{TotalVol}_t$	0.035***	0.035***
	(.004)		*(0.003)*	*(0.003)*
		RV^{NBBO}_t	-0.008***	-0.017***
			(0.003)	*(0.002)*
		Quoted Spread	1.044***	1.047***
			(0.007)	*(0.005)*
		Inverse Mills Ratio	0.060	
			(0.078)	
Entity-Effects	Yes	Yes	Yes	Yes
Day-Effects	Yes	Yes	Yes	Yes
R-squared	55.91		37.70	40.87
Num. Obs	6,490,988		5,559,614	7,106,640

Figure 12: Rule 605 Reported Spreads and Realized Profitability

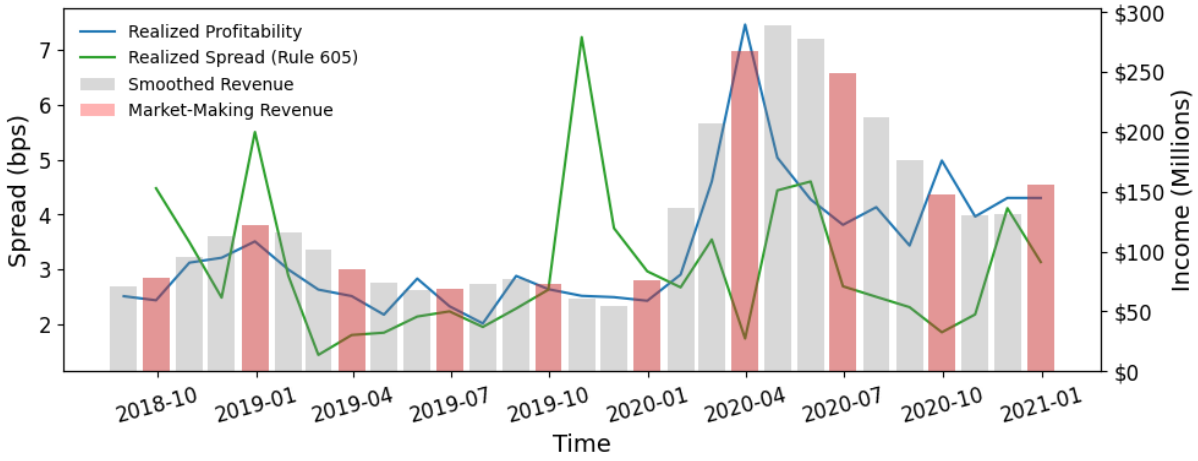
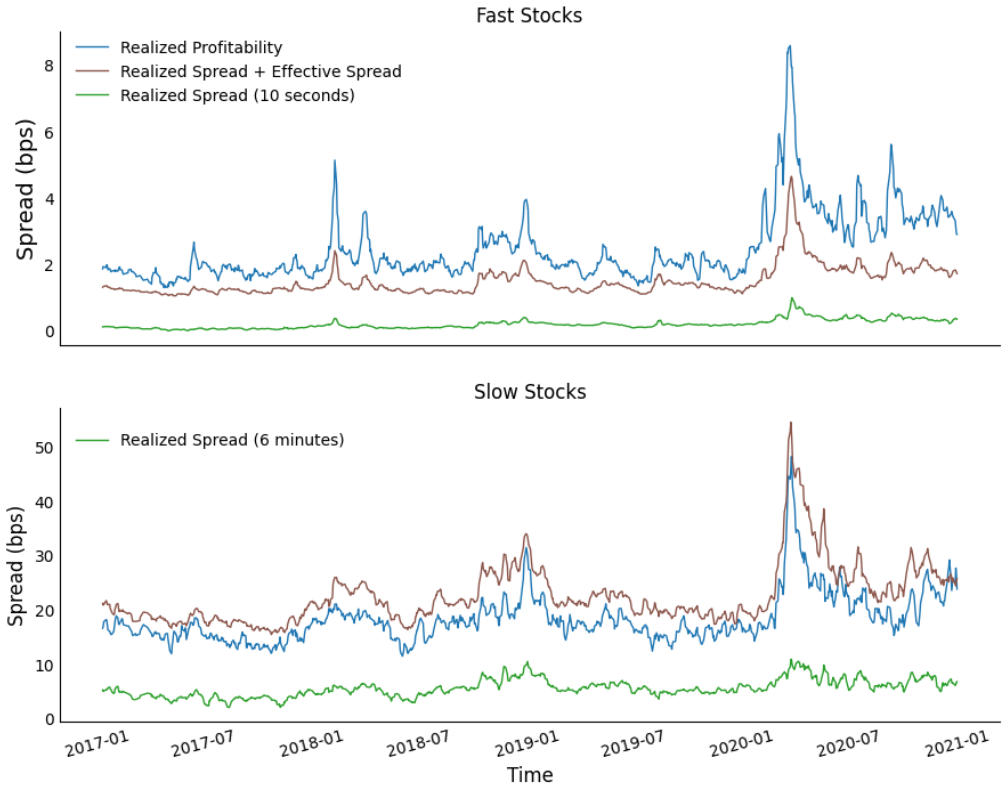


Figure plots the monthly average realized spread (green line), monthly average realized profitability, and Virtu's quarterly market-making revenue (red bar) together with the monthly interpolation of the revenue (grey bar). Average realized spread is computed monthly by aggregating reported realized spreads of a group of matched securities (both in Virtu's Rule 605 and in our sample) using the executed number of shares as weights. Average realized profitability is computed for the same group using our realized profitability data and the same weights. Market-making revenue is the quarterly trading income of Virtu's market-making segment (from Virtu's 10K filings).

Figure 13: Adding Effective Spread to Conventional Realized Spread



The dollar-volume weighted realized profitability rp (blue) for the sample of "fast" securities is plotted in the top panel alongside the dollar-volume weighted realized spread computed with a 10-second horizon with the effective spread (brown) and without (green). The bottom panel plots a similar time series for the "slow" stocks but with the realized spread computed with a 6-minute horizon.

Figure 14: Distribution of τ

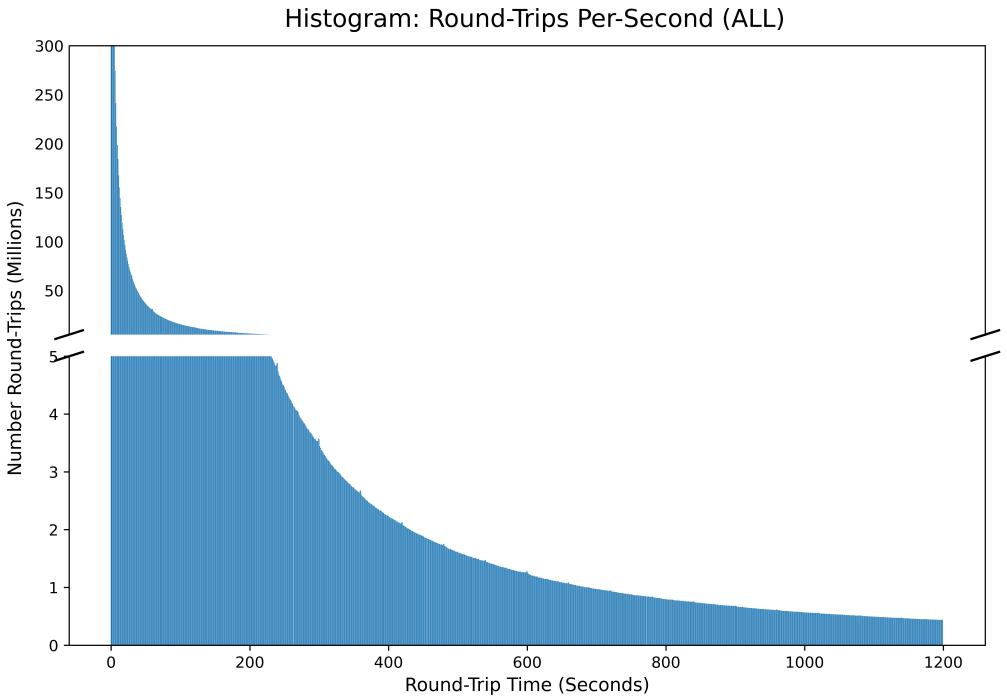


Figure plots the histogram of the round trip time τ (restricted to up to 1200 seconds for visual clarity). The x-axis corresponds to the inventory turnaround time τ, the y-axis the total number of trips that are turned around at τ (x-axis) from their initiation. Using dollar volume instead of the number of trips gives similar distribution.

Figure 15: Deviation of Realized Spreads From Market-making Revenue

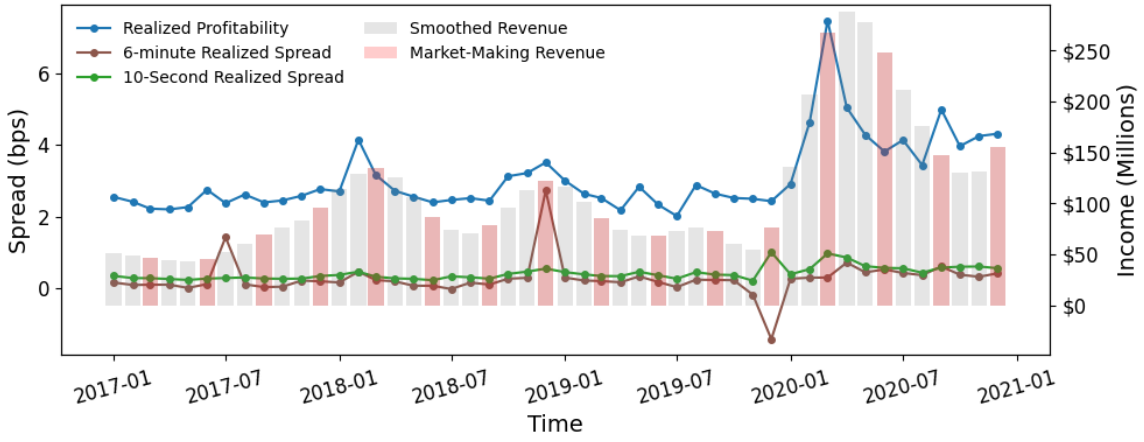
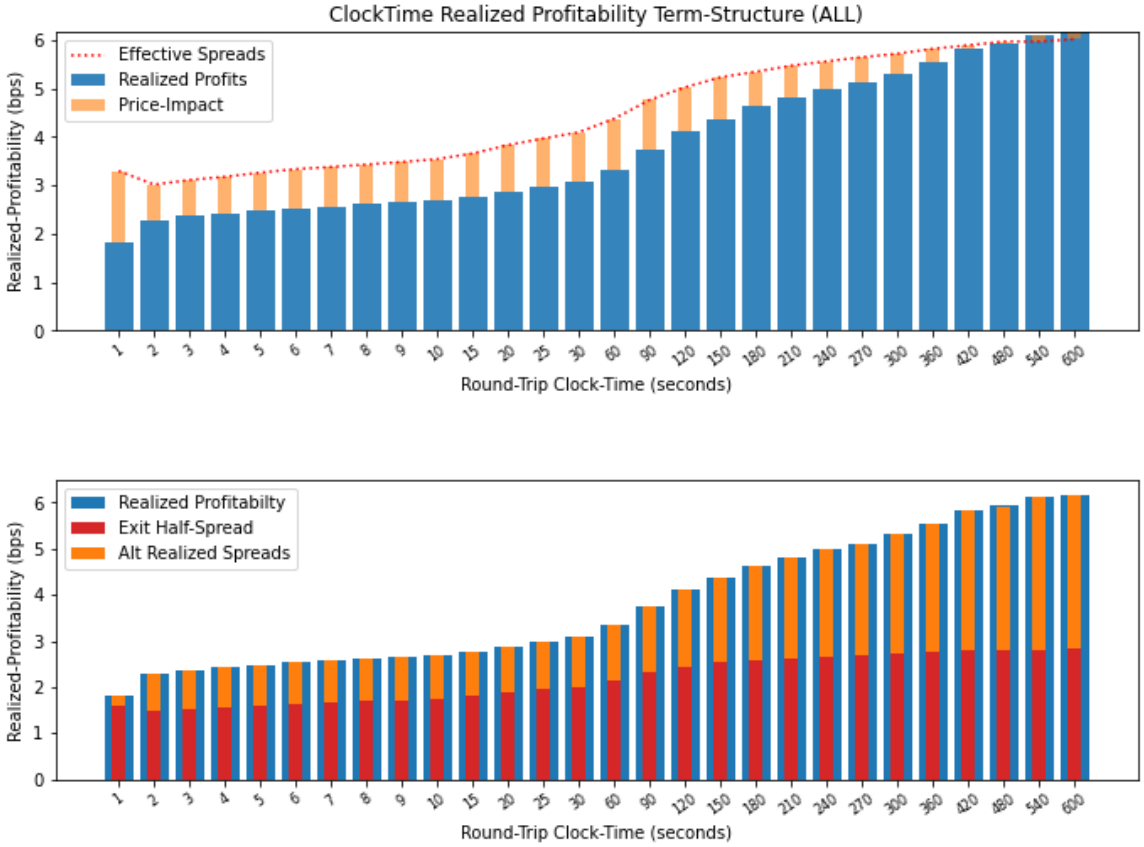


Figure shows the time-series of (dollar-volume-weighted) average realized profitability, average realized spreads under both 10 seconds and 6 minutes, and quarterly market-making revenue (with monthly interpolation). Monthly measures of realized profitability are computed by taking the dollar-volume-weighted average of the realized profitability of all round trips in that month. Monthly measures of realized spread are computed by taking the dollar-volume-weighted average of the realized spread of all trades in that month. Quarterly market-making revenue data comes from Virtu's quarterly financial report: the trading income under the market-making segment.

Figure 16: Aggregate Realized Profitability



The figure plots out the term structure of the realized profitability rp over clock-time horizons. Each bar shows the dollar-volume weighted-average rp (blue bars) of all round trips with a turnaround time between the specified blocks of time. The first two blocks are composed of round trips of 0-1 seconds and 1-2 seconds; the last two blocks are composed of trips that took 480-540 seconds and 540-600 seconds. The top panel decomposes realized profitability into the sum of effective spreads (red-dashed line) and price-impact (orange bars). The bottom panel decomposes rp into the exiting half-spread (red bars) and the alternative realized spread component with endogenous τ (orange bars).

Figure 17: Sharpe Ratio of Liquidity Provision

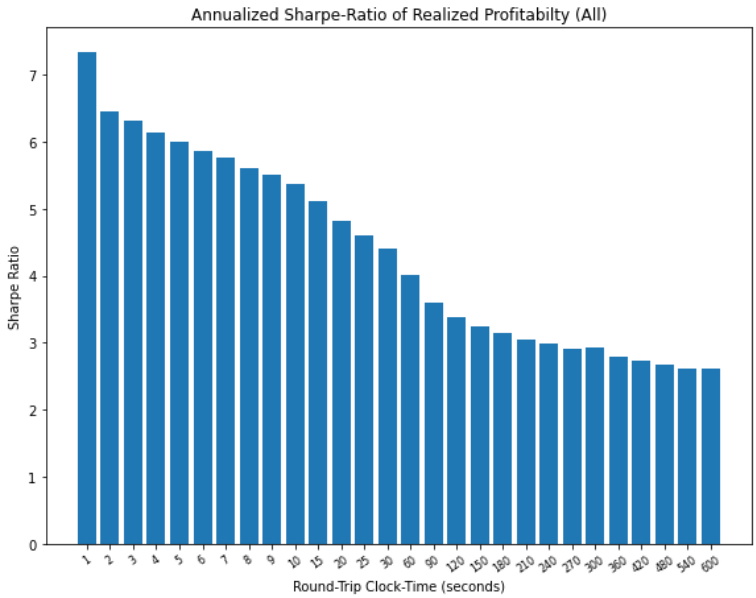
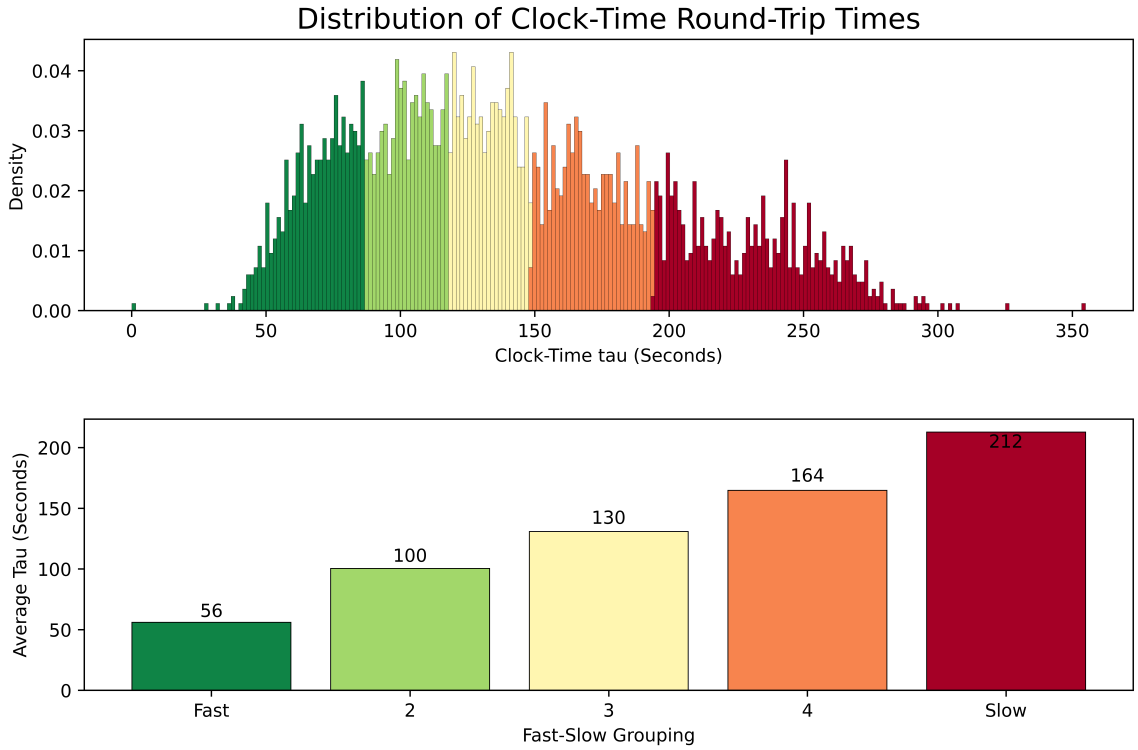
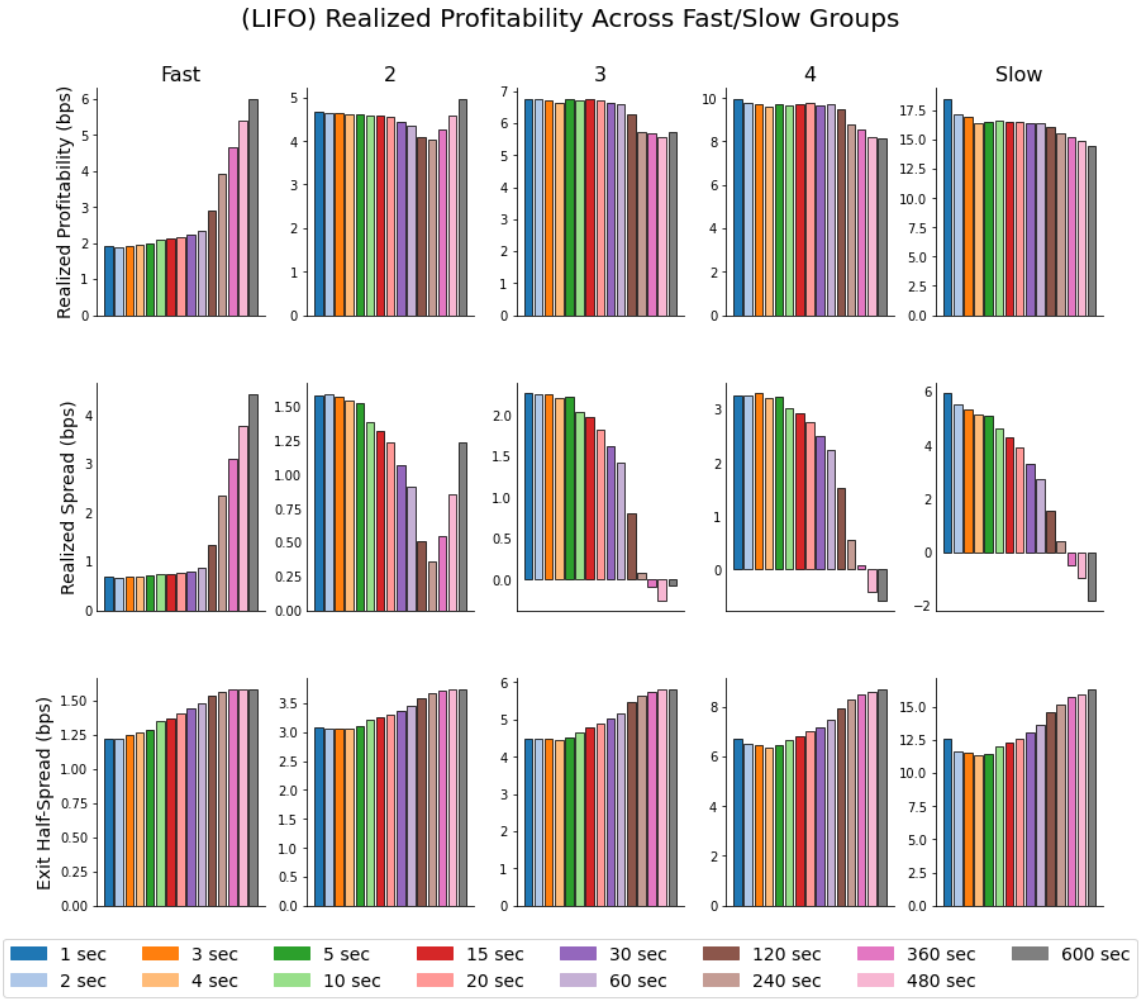


Figure 18: Distribution of τ (Cross-section)



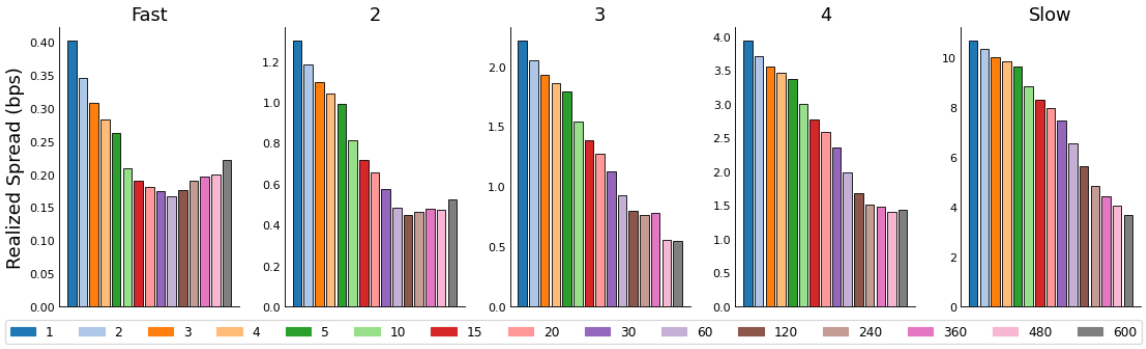
The top panel of this figure plots out the distribution of the average turnaround time τ of the individual stocks in our sample. The y axis denotes the percentage of stocks with an average turnaround time between the range marked by the edges of the bars along the x-axis. The sample is split into quintile grouping based on τ and is color-coated on a fast (green) to slow (red) spectrum. The bottom panel plots out the dollar-volume weighted average turnaround time of each quintile grouping in seconds.

Figure 19: Realized Profitability



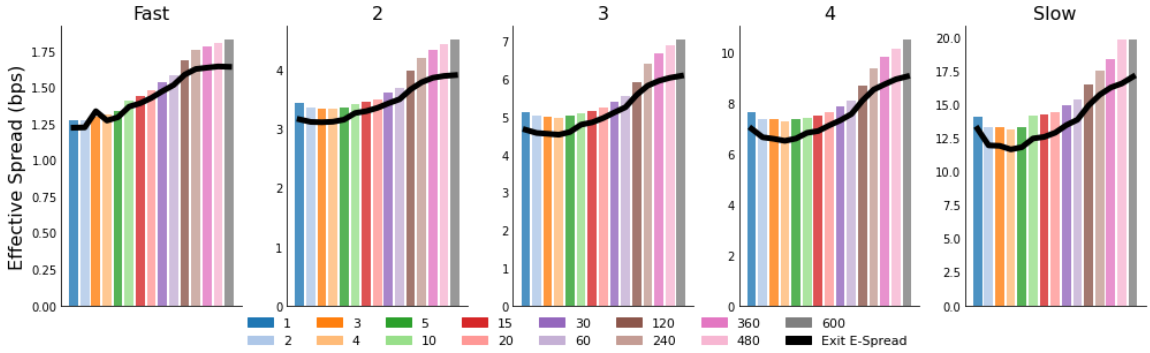
The figure plots out the term structure over clock-time horizons of the realized profitability and its component elements across τ quintile groupings with the fastest securities in the leftmost column and the slowest in the rightmost. The top row shows the term-structure of $rp_{t,\tau}$, the middle row shows the realized spread component $rs_{t,\tau}$, and the final row the exiting effective spread $\delta_t(M_{t+\tau} - P_{t+\tau})$.

Figure 20: Realized Spread Term Structures by Groups



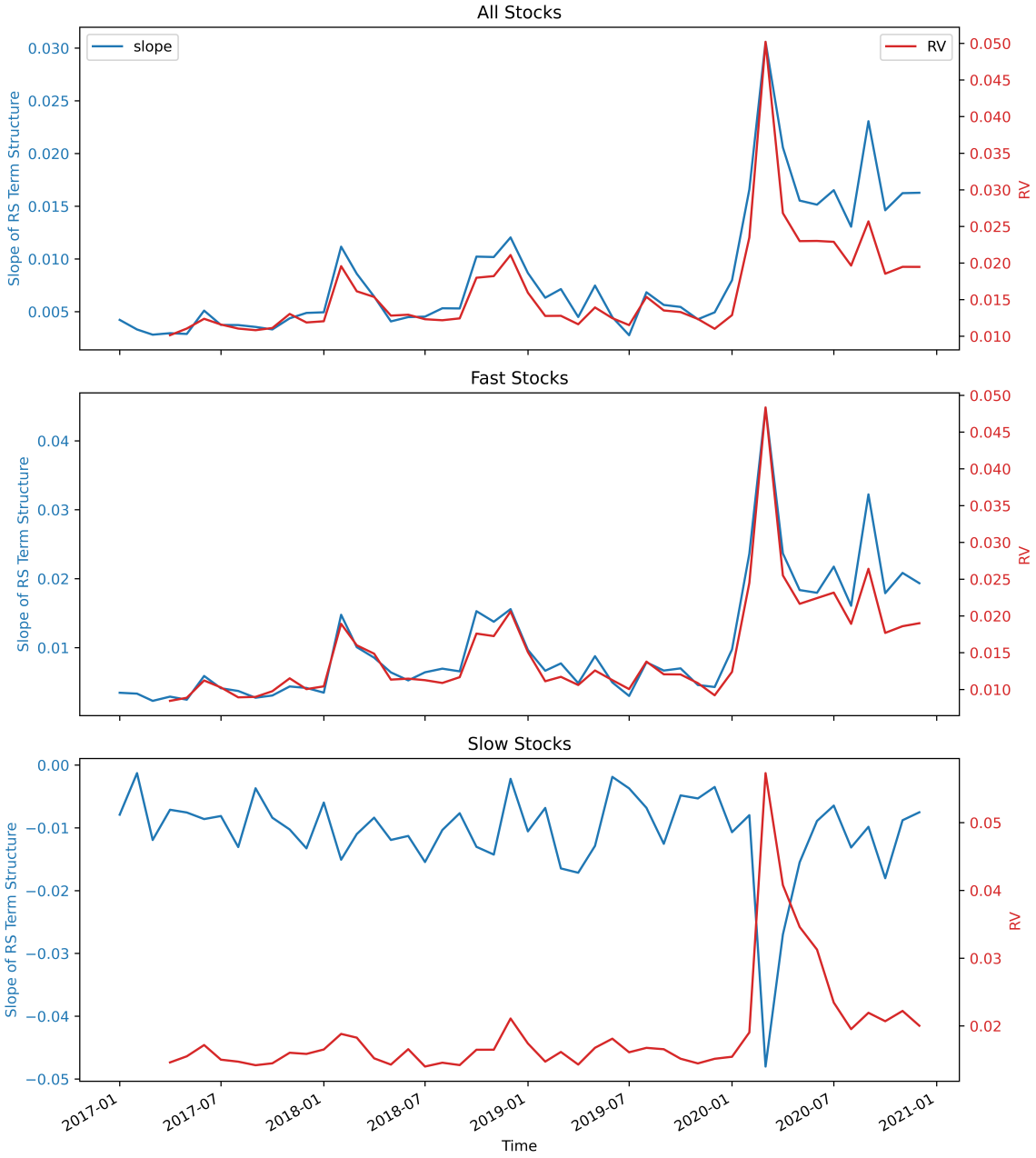
The figure plots the term structure over clock-time horizons of conventionally measured realized spread across τ quintile groupings with the fastest securities in the leftmost column and the slowest in the rightmost.

Figure 21: Effective Spreads by Groups



Here we plot out the term structure of the dollar-volume weighted effective spreads at the beginning of the round-trips at different horizons across τ quintile groupings with the fastest securities in the leftmost column and the slowest in the rightmost. The black solid line outlines the values of the effective spread at the exit of the trips.

Figure 22: Term-Structure Steepness and Volatility



Slope estimates $\hat{\beta}$ from monthly regressions of round-trip profitability onto a constant and turn-around time τ of regression specification: $_{t,\tau} = \alpha + \beta\tau + \epsilon_t$ are plotted alongside the dollar-volume weighted average total realized variation RV for the security subsample over time. The top panel plots out these values for the full sample, the middle panel repeats the exercise for the subset of "fast" stocks while the last panel does so for "slow" stocks.

Figure 23: Realized Profitability Compared with Realized Spreads for Fast Stocks (left) and Slow Stocks (right)

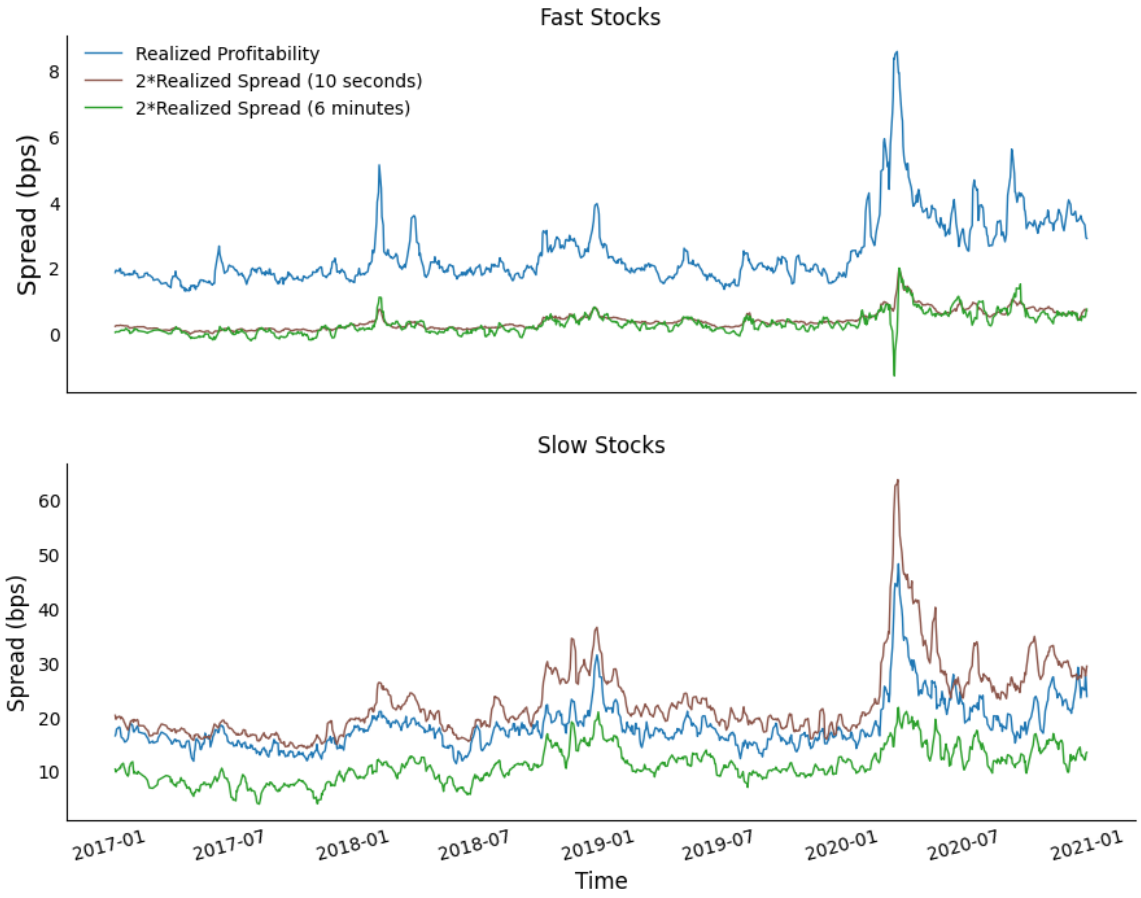


Figure plots the time series of dollar-volume-weighted average realized profitability and dollar-volume-weighted average realized spreads measured with both 10 seconds τ and 6 minutes τ.

Figure 24: Aggregate Realized Profitability (FIFO) across Days Sorted by Order Imbalance

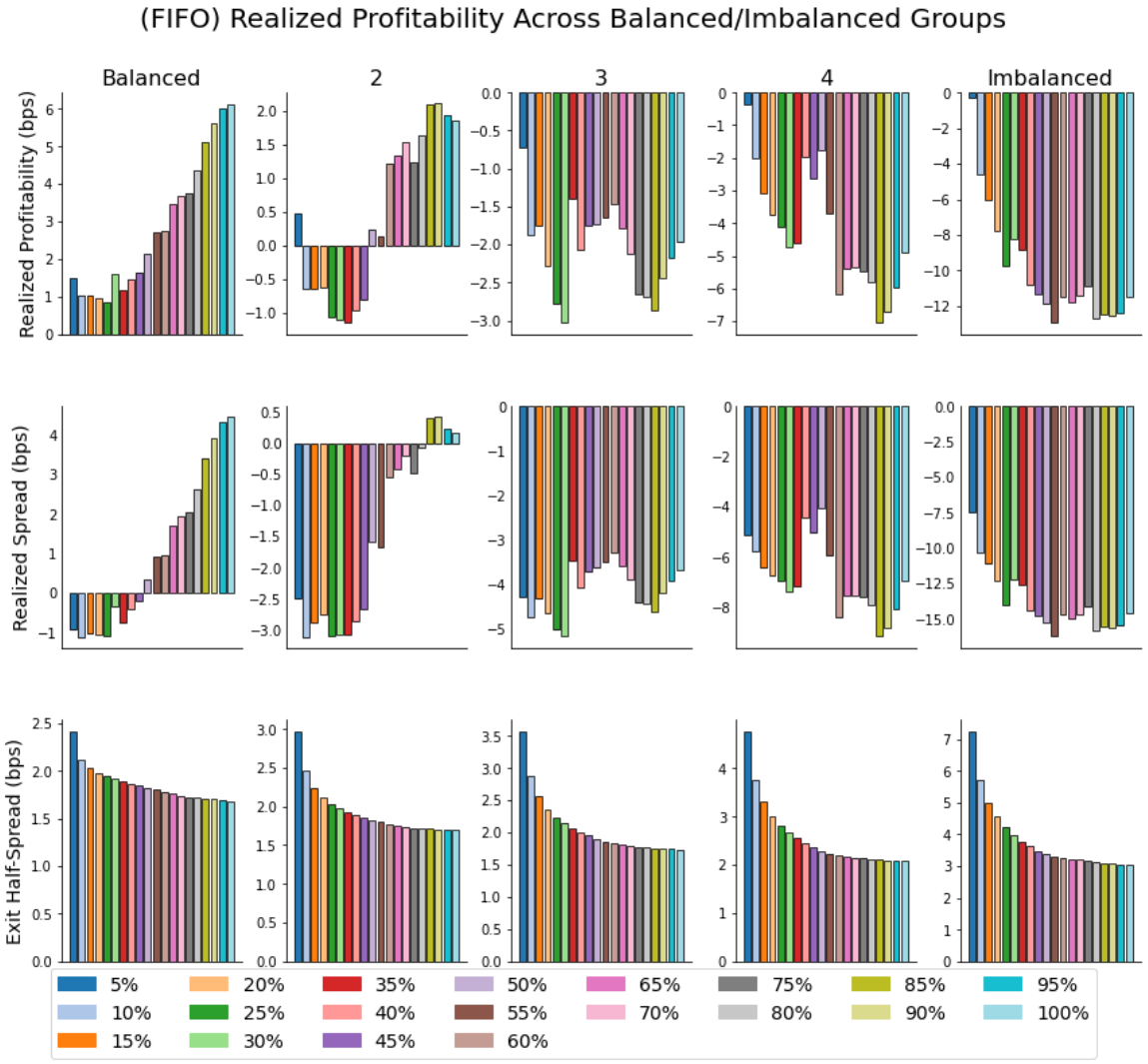
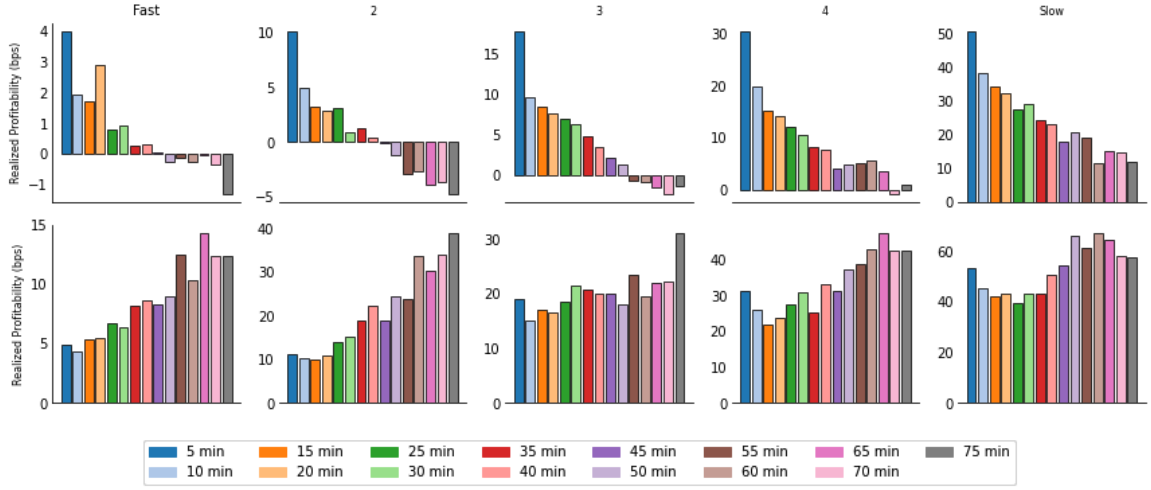


Figure 25: Aggregate Realized Profitability (FIFO) across Fast and Slow Stocks



We use all sample days (top) and sample days with low imbalance days (bottom) to generate the realized profitability term structures for fast/slow stock groupings.

Table 9: Realized profitability and firm characteristics by groups

We sort stocks into decile groups based on their average inventory turnaround time. In Panel A we compute the dollar-volume-weighted average realized spread for each group using round trips of all stocks in that group, similarly, we also compute and report the average Sharpe ratio of the realized profitability, average inventory turnaround time (in trade time and clock time) and average effective spreads. In Panel B we describe the characteristics of stocks in each group. All firm characteristics are measured at the end of each fiscal year and then averaged across stocks using lag firm size as weights. MktCap (size) is price times shares outstanding; investment rate is the % change in total asset; book-to-market is the ratio of book equity to size; gross-profitability is revenues minus cost of goods sold over total asset; ROE is income before extraordinary items over lagged book equity; trading turnover is average daily volume over shares outstanding; Market beta is computed annually using daily returns; idiosyncratic volatility is the standard deviation of the residual from the market model regression.

Panel A: Trade Variables					
	Fast	2	3	4	Slow
Realized profitability	2.45	4.43	6.31	9.11	15.53
Sharpe ratio	3.04	3.30	3.74	4.56	5.74
Entering Effective Spread	1.44	3.68	5.68	8.50	16.18
Exiting Effective Spread	1.39	3.43	5.23	7.73	14.36
Realized-Spread Component	1.07	1.01	1.08	1.38	1.17
τ (in # of trades)	211	60	40	27	17
τ (in seconds)	46	100	131	165	213
Panel B: Other Characteristics					
	Fast	2	3	4	Slow
MktCap (Billions)	61.96	5.71	2.25	1.30	0.70
Investment Rate	0.09	0.13	0.14	0.12	0.12
Book-to-Market	0.37	0.44	0.48	0.50	0.62
Gross-Profitabilty	0.28	0.27	0.26	0.26	0.19
ROE	0.25	0.16	0.10	0.12	0.10
Trading Turnover	6.56	9.33	8.16	6.11	4.61
Market Beta	0.98	1.09	1.11	1.09	0.97
Idiosyncratic Vol	0.01	0.02	0.02	0.02	0.02

Table 10: Average Realized Profitability for Double Sorted Groups

Table reports dollar-volume-weighted average realized profitability for stock groups sorted first by size and then average τ (left), and dollar-volume-weighted average realized profitability for stock groups sorted first by book-to-market and then average τ (right). "Small/Large" corresponds to the size grouping and "Low/High" corresponds to the book-to-market grouping.

	Fast	2	3	4	Slow		Fast	2	3	4	Slow
Small	16.02	17.61	19.62	28.63	34.07	Low	2.76	4.11	5.59	6.74	10.75
2	10.33	9.68	9.98	11.75	17.49	2	1.85	3.80	5.68	7.51	12.45
3	5.74	6.37	6.57	7.31	10.02	3	1.94	3.90	5.82	8.47	14.46
4	4.05	3.69	4.20	4.75	6.40	4	1.59	4.29	7.38	12.02	23.68
Large	2.22	1.64	2.28	2.45	4.05	High	1.92	4.27	7.26	10.24	20.39

REFERENCES

Amihud, Y., "Illiquidity and stock returns: cross-section and time-series effects", Journal of Financial Markets **5**, 1, 31–56 (2002).

Andersen, T. G., T. Bollerslev, F. X. Diebold and H. Ebens, "The distribution of realized stock return volatility", Journal of financial economics **61**, 1, 43–76 (2001).

Bennett, P. and L. Wei, "Market structure, fragmentation, and market quality", Journal of Financial Markets **9**, 1, 49–78 (2006).

Bessembinder, H., "Quote-based competition and trade execution costs in nyse-listed stocks", Journal of Financial Economics **70**, 3, 385–422 (2003).

Bessembinder, H. and H. M. Kaufman, "A comparison of trade execution costs for nyse and nasdaq-listed stocks", The Journal of Financial and Quantitative Analysis **32**, 3, 287–310 (1997).

Boehmer, B. and E. Boehmer, "Trading your neighbor's etfs: Competition or fragmentation?", Journal of Banking & Finance **27**, 9, 1667–1703 (2003).

Chakravarty, S., P. Jain, J. Upson and R. Wood, "Clean sweep: Informed trading through intermarket sweep orders", Journal of Financial and Quantitative Analysis pp. 415–435 (2012).

Chowdhry, B. and V. Nanda, "Stabilization, syndication, and pricing of ipos", Journal of Financial and Quantitative Analysis pp. 25–42 (1996).

Comerton-Forde, C., T. Hendershott, C. M. Jones, P. C. Moulton and M. S. Seasholes, "Time variation in liquidity: The role of market-maker inventories and revenues", The journal of finance **65**, 1, 295–331 (2010).

Conrad, J. and S. Wahal, "The term structure of liquidity provision", Journal of Financial Economics **136**, 1, 239–259 (2020).

Demsetz, H., "The cost of transacting", The quarterly journal of economics **82**, 1, 33–53 (1968).

Easley, D., M. M. L. De Prado and M. O'Hara, "The microstructure of the "flash crash": flow toxicity, liquidity crashes, and the probability of informed trading", The Journal of Portfolio Management **37**, 2, 118–128 (2011).

Easley, D., N. M. Kiefer, M. O'Hara and J. B. Paperman, "Liquidity, information, and infrequently traded stocks", The Journal of Finance **51**, 4, 1405–1436 (1996).

Economides, N., "The economics of networks", International journal of industrial organization **14**, 6, 673–699 (1996).

Foucault, T., O. Kadan and E. Kandel, "Limit Order Book as a Market for Liquidity", The Review of Financial Studies **18**, 4, 1171–1217 (2005).

Glosten, L. R., "Is the electronic open limit order book inevitable?", The Journal of Finance **49**, 4, 1127–1161 (1994).

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

Glosten, L. R. and P. R. Milgrom, "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders", Journal of Financial Economics **14**, 1, 71–100 (1985).

Harris, L., "The winners and losers of the zero-sum game: The origins of trading profits, price efficiency and market liquidity", (1993).

Hasbrouck, J., "Trades, quotes, inventories, and information", Journal of Financial Economics **22**, 2, 229–252 (1988).

Hasbrouck, J., "One security, many markets: Determining the contributions to price discovery", The journal of Finance **50**, 4, 1175–1199 (1995).

Hasbrouck, J. and G. Sofianos, "The trades of market makers: An empirical analysis of nyse specialists", The Journal of Finance **48**, 5, 1565–1593 (1993).

Haslag, P. H. and M. C. Ringgenberg, "The demise of the nyse and nasdaq: Market quality in the age of market fragmentation", in "Fourth Annual Conference on Financial Market Regulation", (2020).

Heckman, J. J., "Sample selection bias as a specification error", Econometrica: Journal of the econometric society pp. 153–161 (1979).

Ho, T. and H. R. Stoll, "Optimal dealer pricing under transactions and return uncertainty", Journal of Financial Economics **9**, 1, 47–73 (1981).

Ho, T. S. Y. and H. R. Stoll, "The dynamics of dealer markets under competition", The Journal of Finance **38**, 4, 1053–1074 (1983).

Holden, C. W. and S. Jacobsen, "Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions", The Journal of Finance **69**, 4, 1747–1785 (2014).

Huang, R. D. and H. R. Stoll, "Market microstructure and stock return predictions", The Review of Financial Studies **7**, 1, 179–213 (1994).

Huang, R. D. and H. R. Stoll, "Dealer versus auction markets: A paired comparison of execution costs on nasdaq and the nyse", Journal of Financial economics **41**, 3, 313–357 (1996).

Kraus, A. and H. R. Stoll, "Price impacts of block trading on the new york stock exchange", The Journal of Finance **27**, 3, 569–588 (1972).

Kyle, A. S., "Informed Speculation with Imperfect Competition", The Review of Economic Studies **56**, 3, 317–355 (1989).

Lee, C. M. C. and M. J. Ready, "Inferring trade direction from intraday data", The Journal of Finance **46**, 2, 733–746 (1991).

Madhavan, A., "Consolidation, fragmentation, and the disclosure of trading information", The Review of Financial Studies **8**, 3, 579–603 (1995).

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

O'Hara, M. and M. Ye, "Is market fragmentation harming market quality?", Journal of Financial Economics **100**, 3, 459–474 (2011).

Pagano, M., "Trading volume and asset liquidity", The Quarterly Journal of Economics **104**, 2, 255–274 (1989).

Stoll, H. R., "The supply of dealer services in securities markets", The Journal of Finance **33**, 4, 1133–1151 (1978).

Stoll, H. R., "Inferring the components of the bid-ask spread: Theory and empirical tests", The Journal of Finance **44**, 1, 115–134 (1989).

Zhang, L., P. A. Mykland and Y. Aït-Sahalia, "A tale of two time scales: Determining integrated volatility with noisy high-frequency data", Journal of the American Statistical Association **100**, 472, 1394–1411 (2005).

APPENDIX A

APPENDIX TO CHAPTER 1

ISO Ordering

Per the SEC's description of ISOs:

Simultaneously with the routing of the limit order identified as an intermarket sweep order, one or more additional limit orders, as necessary, are routed to execute against the full displayed size of any protected bid, in the case of a limit order to sell, or the full displayed size of any protected offer, in the case of a limit order to buy, for the NMS stock with a price that is superior to the limit price of the limit order identified as an intermarket sweep order. These additional routed orders also must be marked as intermarket sweep orders.

The language of the rule calls only for the "simultaneous" routing of order. This is of course impossible in the most literal sense of the word and ISO orders are presented with different time-stamps in the data. The language concerning the order of ISO submission remained opaque in a 2008 SEC memorandum for Rule 611 FAQs. One section reads:

"... whenever an order-router intends to sweep one or more inferior prices, an ISO must be routed to execute against every better-priced protected quotation..."

One would think that by first sending ISOs to better-priced exchanges they may assure compliance with the regulation should their trading be interrupted for whatever reason. Submitting ISOs per this ordering would go a long way towards justifying the methodology. This, I stress, is just conjecture. In another section the memorandum reads:

"... To meet this requirement, the broker-dealer will need to utilize an automated system that is capable of ascertaining current protected quotations and simultaneously routing the necessary ISOs. ..."

The language calling for simultaneous routing mirrors that found within the letter of Rule 611, but what is meant by simultaneity remains unclear. Absent clarification and no obvious reasoning as to why ISOs shouldn't be routed best-price-first I believe it is reasonable to assume a random ordering. The best-price-first ordering the methodology lines up perfectly with the counterfactual trade measurement, with random ordering the methodology produces a noisy but not biased estimate, and with best-price-last ordering the methodology producing an upper bound for these costs.

RV Measurements

For each day d, for each exchange E, I observe a series of price observations, denoted P_t. These could be either quoted or transaction prices. It is assumed the series of prices $\{P_t\}_{d,E}$ are noisy observations of the "true" price X_t in the sense that the observed log price is equal to the "true" latent log price plus some noise, η_t:

$$\ln P_t = \ln X_t + \eta_t$$

The stochastic microstructure noise η_t is assumed to be centered at zero with a constant within-day volatility but may be heteroskedastic across days. Given the contamination of the observed prices due to microstructure noise, the usual estimator of the process's quadratic variation (shown below) would be inappropriate:

$$[\ln P, \ln P]_{\Pi_{E,d}^{(all)}} = \sum_{i=1}^{n_d} (lnP_{t_{d,i}} - lnP_{t_{d,i-1}})^2 \quad \text{where:} \;\; n_d = |\Pi_{E,d}^{(all)}|$$

Where $\Pi_{E,d}^{(all)} = \{t_{d,0}, t_{d,1}, t_{d,2}, ..., t_{d,n_d}\}$ denotes the whole set of available sampling points for $\{P_t\}_{d,E}$ on day d with $t_{d,0} < t_{d,1} < ... < t_{d,n_d}$. Due to the presence of measurement error it is well known that as the set of sampling points gets large the expected value of the above estimator diverges, $\lim_{n_d \to \infty} \langle [lnP, lnP]_{\Pi^{(all)}} \rangle = \infty$. In order to address this issue I employ the methodology set out in Zhang *et al.* (2005) which combines a bias-correction with averaging in order to make use of all available data. Averaging works to make use of the whole data by sampling at predetermined frequency across non-intersecting sampling grids $\Pi \subset \Pi^{(all)}$.

Under the assumption that the microstructure noise is independently identically distributed and orthogonal to the efficient price process, the expected value of the realized variation from using a sub-partition $\Pi \subset \Pi_{d,E}^{(all)}$ would be given by:

$$\langle [lnP, lnP]_{\Pi} \rangle = [lnX, lnX]_{\Pi} + 2n\gamma_{\eta}(0) \tag{A.1}$$

Where $\gamma_{\eta}(0)$ is the variance of the microstructure noise. Note that as the sample size n grows large, the bias term $2n\gamma_{\eta}(0)$ in equation (A.1) dominates in expectation and $\langle [lnP, lnP]_{\Pi} \rangle \to \infty$ as $n \to \infty$. The good news is that if $|\Pi_{d,E}^{(all)}| = n_d$ is large enough, then the microstructure noise variance could be proxied by:

$$\frac{[lnP, lnP]_{\Pi_d^{(all)}}}{2n_d} \approx \gamma_{\eta}(0) \tag{A.2}$$

So subtracting $\frac{n}{n_d}[lnP, lnP]_{\Pi_d^{(all)}}$ from $[lnP, lnP]_{\Pi}$ would result in a consistent estimator; this is the bias correction.

In order to implement averaging, for a sampling frequency K, define the j^{th} partition $\Pi_K(j) \subset \Pi_d^{(all)}$ as

$$\Pi_K(j) = \{t_{Ki+j} | i \in \{0, 1, 2, ..., \lfloor \frac{n_d - 1 - j}{K} \rfloor\}\}$$

So for a set integer value K we can define the "averaged" realized variance as:

$$\overline{[lnP, lnP]}_K = \frac{1}{K} \sum_{j=0}^{K-1} [lnP, lnP]_{\Pi_K(j)}$$

The averaged bias corrected (ZMA) estimator for the d^{th} day (where $\bar{n}_k = \frac{1}{K} \sum_{j=0}^{K-1} \lfloor \frac{n_d - 1 - j}{K} \rfloor$ is the average cardinality of the subpartitions) as:

$$\widehat{RV}_{d,K}^{E} = \overline{[lnP, lnP]}_K - \frac{\bar{n}_K}{n_d} [lnP, lnP]_{\Pi_d^{(all)}} \tag{A.3}$$

It is typically assumed that η_t is orthogonal to the true log price and exhibits no autocorrelation, so $\langle \eta_s \eta_t \rangle = 0$ if $s \neq t$ and $\langle \ln X_t \eta_t \rangle = 0$. Under these conditions the ZMA estimator would be consistent, in theory it would not be free of bias in if these assumptions are not true. However one may show that any bias resulting from the violation of those assumptions would must lie between ξ and $-\xi$ where $\xi > 0$ is a quantity proportional to the level of microstructure noise variance $\gamma_\eta(0)$.

The sampling frequencies K used for the averaged estimators are chosen so that the sampling points in each subpartition are approximately 300 seconds (5 minutes) apart. On average, sampling points were 301.81 seconds apart with a standard deviation of 0.91 seconds, so I feel that my estimates should be consistent with other literature which follow the standard recommendation for highly liquid assets of Andersen *et al.* (2001) to sample every five minutes.

Demonstrative Example:

Consider the case when we have 108,000 price observations within a single day (five observations per second) and we are interested in sampling every five minutes. Given that there are 72 non-overlapping five minute intervals for the 6-hour sample period, and if the observation times are roughly equally spaced, then the 1^{st} and $\frac{108{,}000}{72} = 1{,}500^{\text{th}}$ observation would be roughly 300 seconds apart. In this case the averaged RV estimator ($K = 1500$), $\overline{[lnP, lnP]}_{1500}$ would be the average of the following 1500 individual non-overlapping RVs:

$$
\begin{aligned}
[lnP, lnP]_{\Pi_{1500}(0)} &= \sum_{i=0}^{71} (\ln P_{(i+1)1500+0} - \ln P_{(i)1500+0})^2, \\
[lnP, lnP]_{\Pi_{1500}(1)} &= \sum_{i=0}^{71} (\ln P_{(i+1)1500+1} - \ln P_{(i)1500+1})^2, \\
[lnP, lnP]_{\Pi_{1500}(2)} &= \sum_{i=0}^{71} (\ln P_{(i+1)1500+2} - \ln P_{(i)1500+2})^2, \\
&\vdots
\end{aligned}
$$

With $\bar{n}_{1500} = 71$ the bias correction term would be $\frac{71}{10800}$ times the total RV ($[lnP, lnP]_{\Pi_d^{(all)}}$) using all 108,000 observations.

Table A.1: Gross ISO Cost Results Across Subsamples

Regression results for the log-log panel regressions of the form:

$$\ln ISOcps_{i,t} = \alpha + \beta_1' \ln F_{i,t} + \beta_2' \ln X_{i,t} + \epsilon_{i,t}$$

for all securities, the top 50 securities, and those securities which are traded on 7 or more exchanges. Fragmentation measures are comprised of $F_{i,t} = \left[\text{Disp}[RV]_{i,t}, AOS_{i,t}, (1 - HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}\right]$. Control variables captured in the design vector X are $\{RV^{nbbo}, Spread, sweepProp, ILLIQ\}$. The first column for each sample reports the regression results excluding the control variable X from the specification. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Month-time and security fixed effects are included in all specifications.

Variable	Gross ISO Costs Per Share					
	All		≤ 3 Trade Venues		> 9 Trade Venues	
Constant	3.643***	3.494***	2.913***	4.798***	5.022***	7.009***
	(0.061)	*(0.078)*	*(0.078)*	*(0.108)*	*(0.169)*	*(0.198)*
Disp[*RV*]	0.177***	0.165***	0.114***	0.062***	0.257***	0.362***
	(0.004)	*(0.005)*	*(0.005)*	*(0.006)*	*(0.010)*	*(0.014)*
AOS	−0.003	−0.041***	0.050***	−0.010	−0.052***	−0.066
	(0.004)	*(0.003)*	*(0.009)*	*(0.007)*	*(0.008)*	*(0.007)*
$(1 - HHI)$	0.868***	1.001***	0.482***	0.640***	1.679***	1.676***
	(0.039)	*(0.038)*	*(0.048)*	*(0.047)*	*(0.161)*	*(0.150)*
$\frac{DarkVol}{TotalVol}$	0.087***	0.088***	0.082***	0.081***	0.084***	0.107***
	(0.006)	*(0.006)*	*(0.010)*	*(0.011)*	*(0.020)*	*(0.018)*
RV^{NBBO}		0.010**		0.050***		−0.142***
		(0.005)		*(0.008)*		*(0.014)*
QuotedSpread		0.379***		0.477***		0.316***
		(0.011)		*(0.019)*		*(0.023)*
SweepProportion		0.089***		0.116***		0.040**
		(0.007)		*(0.010)*		*(0.019)*
ILLIQ		0.017***		0.018***		0.012***
		(0.002)		*(0.004)*		*(0.002)*
Entity-Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month-Effects	Yes	Yes	Yes	Yes	Yes	Yes
R2	41.54	37.52	27.50	27.39	69.93	73.14
Num. Obs	1,10,530		132,266		114,100	
Num. Entities	2,637		584		238	

Table A.2: Gross ISO Costs Results Across Small/Large Subsamples

Regression results for the log-log panel regressions of the form:

$$\ln ISOcps_{i,t} = \alpha + \beta_1' \ln F_{i,t} + \beta_2' \ln X_{i,t} + \epsilon_{i,t}$$

for all securities and the market-cap size quintile subsets. Fragmentation measures are comprised of $F_{i,t} = \left[\text{Disp}[RV]_{i,t}, AOS_{i,t}, (1-HHI_{i,t}), \frac{DarkVol}{TotalVol}_{i,t}\right]$. Control variables captured in the design vector X are $\{RV^{nbbo},\ Spread,\ sweepProp,\ ILLIQ\}$. Significance is denoted at the 10% *, 5% **, and 1% *** levels which are determined based on the entity-month clustered standard errors reported in parenthesis below the coefficients. Month-time and security fixed effects are included in all specifications.

Variable	Gross ISO Costs Per Share					
	Smallest 20%		Middle 60%		Largest 20%	
Constant	4.050^{***}	6.261^{***}	3.612^{***}	5.397^{***}	3.465^{***}	4.707^{***}
	(0.106)	*(0.122)*	*(0.069)*	*(0.087)*	*(0.123)*	*(0.132)*
Disp[*RV*]	0.167^{***}	0.201^{***}	0.172^{***}	0.151^{***}	0.193^{***}	0.148^{***}
	(0.007)	*(0.010)*	*(0.004)*	*(0.006)*	*(0.007)*	*(0.010)*
AOS	0.015	-0.027^{***}	-0.008^{*}	-0.046^{***}	-0.001	-0.032^{***}
	(0.011)	*(0.001)*	*(0.004)*	*(0.004)*	*(0.005)*	*(0.005)*
$(1 - HHI)$	0.655^{***}	0.752^{***}	0.882^{***}	1.023^{***}	0.828^{***}	0.977^{***}
	(0.062)	*(0.059)*	*(0.044)*	*(0.042)*	*(0.084)*	*(0.088)*
$\frac{DarkVol}{TotalVol}$	0.083^{***}	0.151^{***}	0.081^{***}	0.078^{***}	0.126^{***}	0.061^{***}
	(0.011)	*(0.013)*	*(0.006)*	*(0.006)*	*(0.013)*	*(0.012)*
RV^{NBBO}		-0.070^{***}		0.027^{***}		0.036
		(0.010)		*(0.006)*		*(0.009)*
QuotedSpread		0.292^{***}		0.365^{***}		0.467^{***}
		(0.018)		*(0.012)*		*(0.022)*
SweepProportion		0.126^{***}		0.092^{***}		0.024^{**}
		(0.015)		*(0.008)*		*(0.012)*
ILLIQ		0.045^{***}		0.016^{***}		-0.001
		(0.003)		*(0.002)*		*(0.002)*
Entity-Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month-Effects	Yes	Yes	Yes	Yes	Yes	Yes
R2	33.31	35.91	28.64	26.79	31.26	33.81
Num. Obs	96,963		521,735		185,553	
Num. Entities	388		1,163		388	

APPENDIX B

APPENDIX TO CHAPTER 2

LIFO and FIFO

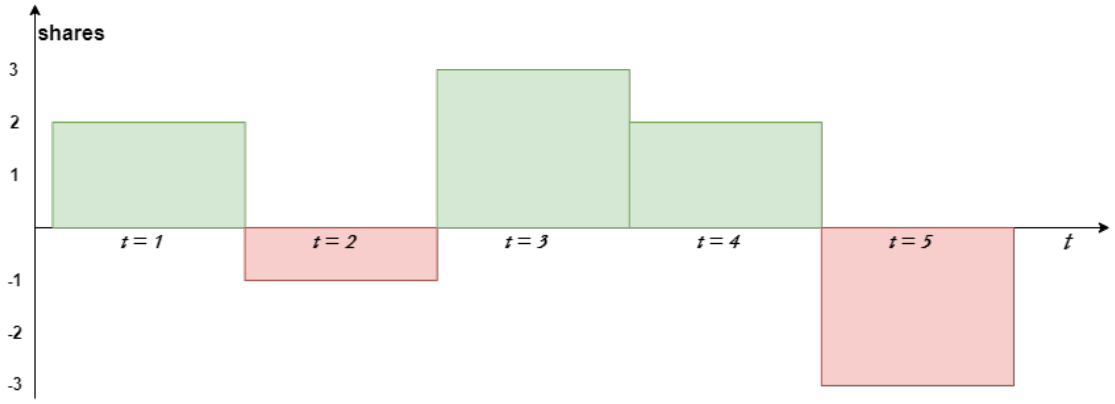
Under the LIFO system each inventory reversing trade is matched to the newest inventory entries in a sequential manner whereas under FIFO the offsetting trade is matched to the oldest pieces of inventory. The difference between the entry and exit time of each inventory entry will be the round trip time (τ) for the trade that initiated that inventory. Figure B.1 illustrates the identification of round trips under the LIFO and FIFO methods using a simple example.

Over any period of time in which all of the LP's inventory positions are completely turned around, the average round trip time τ and dollar-volume-weighted average realized profitability under LIFO would be exactly equal to that under FIFO. However, whenever the LP does not fully turn around it's inventory position, LIFO estimates will deviate from that of FIFO. This discrepancy results from the difference in the set of matched trades between LIFO and FIFO. To illustrate the difference in the selection of trades between these two tracking systems during days with order imbalance, we use an extremely simplified example as shown in Figure B.2. As in the figure, during days with large order imbalance, LIFO matches offsetting trades that are temporally closer to each other than FIFO: average turnaround time is 2 hours under LIFO $((1+3)/2)$ and 5 hours under FIFO $((5+5)/2)$.[1] This feature of LIFO is economically appealing: market makers are more likely to provide liquidity when trades can be turned around faster; given this preference and rational expectations (about the expected time to turnaround a trade) it is reasonable to expect that round trips matched under LIFO were more likely executed by market makers than ULPs.

The advantage of LIFO over FIFO is especially prominent during days when there is large order imbalance. In Table B.1 we sort our stock days into decile groups based on the daily order imbalance level and then report the average round trip time and realized profitability of all round trips from each group. Under LIFO, the average τ increases from 62 seconds for the days with the lowest level of order imbalance to 116 for the days with the highest level of order imbalance. Under FIFO, average turnaround times are much larger and also very sensitive to order imbalance (it increase from 711 seconds for low imbalance day to 5808 seconds for high imbalance day). In terms of realized profitability, the estimates are very close—around 2.7bps—under both LIFO and FIFO during days with small order imbalances (the first two decile groups). However, as we move towards large order imbalance stock days, realized profitability increases gradually to 4.85 bps for the 9th decile group and jump to 8.26 for the group with the largest order imbalance. By contrast, it drops to a dramatic -18.18 bps for the group with the highest order imbalance. Compared to FIFO, LIFO produces much more reasonable estimates of τ across days with and without large order imbalances. Consider the fact that market makers—especially high frequency ones—are extremely averse to holding inventory, we believe the high sensitivity of FIFO estimates to daily order imbalance results from FIFO's tendency to capture trades by ULPs: trades that took extremely long to turnaround were most likely intermediated by long term investors rather than market makers; during days with large order imbalance, FIFO disproportionately select these trades because it

[1] Note that in our analysis we only keep trades that are turned around within a day: during days with order imbalance, the trades, or circles that are not connected by pink dashed lines are omitted from our sample.

Figure B.1: Tracking Round Trips (LIFO and FIFO)



	Trades:	Inventory book:	Round trips (LIFO):	Round trips (FIFO):
t = 1	+2 shares	entry at 1, book: (+2, t = 1)	none	none
t = 2	-1 shares	exit at 2, book: (+1, t = 1)	Trip1: stime = 1, etime = 2 tau = 1, eQ = -1	Trip1: stime = 1, etime = 2 tau = 1, eQ = -1
t = 3	+3 shares	entry at 3, book: (+1, t = 1) (+3, t = 3)	none	none
t = 4	+2 shares	entry at 4, book: (+1, t = 1) (+3, t = 3) (+2, t = 4)	none	none
t = 5	-3 shares	exit at 5, book: (+1, t = 1)	Trip2: stime = 4, etime = 5 tau = 1, eQ = -2	Trip2: stime = 1, etime = 5 tau = 5, eQ = -1
		(+2, t = 3)	Trip3: stime = 3, etime = 5 tau = 2, eQ = -1	Trip3: stime = 3, etime = 5 tau = 2, eQ = -2

Figure illustrates the identification of round trips (how to match off-setting trades to form a round trip) under LIFO and FIFO using an example with 5 trades over a 5-minute window. Light green bars denote market sell orders—or equivalently the LP's buy orders—with sizes shown on the y-axis; light red bars denote the LP's sales similarly. The inventory book records entries of inventory positions with information on the size, direction as well as time of the entry. Under LIFO, off-setting trades are matched with the newest inventory to form a round trip: e.g., at $t = 5$, part of the market buy order is matched with the newest inventory, the 2 shares acquired at $t = 4$, to form the round trip Trip2, which has a turnaround time of 1 minute (exit time 5 − entry time 4). Under FIFO, off-setting trades are matched with the oldest inventory: e.g., at $t = 5$, part of the market buy order is matched with the oldest inventory, the 1 share acquired at $t = 1$, to form the round trip Trip2, which has a turnaround time of 5 minute (exit time 5 − entry time 1).

matches offsetting trades with the oldest inventory.

Aside from being economically meaningful, LIFO also produce estimates of realized profitability that are statistically more robust to order imbalances than FIFO. In the following section we demonstrate how FIFO can introduce mechanical bias in the estimates of realized profitability across market making horizons when there exists large order imbalance.

Figure B.2: Matched Trades under LIFO (left) and FIFO (right)

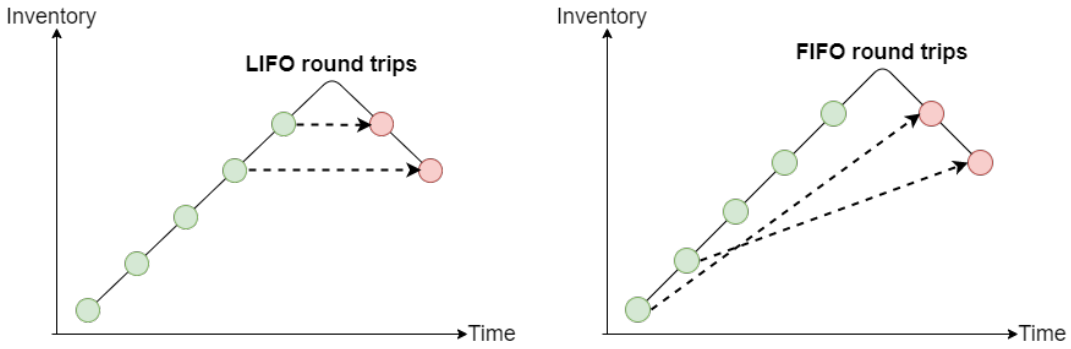


Figure compares the set of trades matched to form round trips under LIFO with the set matched under FIFO in a hypothetical day with 7 trades and order imbalance. The solid black line tracks the cumulative inventory level of the LP throughout the day as shown on the y-axis; light green balls denote market sells (the LP's buys) and light red balls denote market buys, all of unit size and evenly distributed across time (elapsed time between consecutive trades is one hour). The dashed black lines with arrows connect trades to form round trips under LIFO on the left and FIFO on the right. Unmatched trades—the first trades under LIFO and the 3^{rd} through 5^{th} trades under FIFO—show up in the end-of-day inventory as order imbalance.

Table B.1: Average Inventory Turnaround Time and Realized Profitability by Order Imbalance

Table reports average inventory turnaround time τ and realized profitability rp for decile groups of stock days sorted by order imbalance. The sorting variable order imbalance is computed for each stock day as the total order imbalance scaled by the total trading volume $\frac{|\$Buy - \$Sell|}{|\$Buy + \$Sell|}$. For each decile group of stock days, we compute the average τ as the dollar-volume-weighted average τ of all round trips from that group, and the average rp as the dollar-volume-weighted average rp using all round trips from the same group. First row reports average τ using round trips matched under FIFO and second row reports average τ using round trips matched under LIFO. Similarly, the third and fourth row report average rp using round trips matched under FIFO and LIFO respectively. The last row reports the dollar-volume-weighted average value of the sorting variable for each group. Column "All" reports full sample averages (dollar-volume-weighted).

Decile	1	2	3	4	5	6	7	8	9	10	All
FIFO τ (s)	711	875	1,110	1,436	1,801	2,236	2,740	3,368	4,303	5,808	1,567
LIFO τ (s)	62	62	63	64	65	68	72	78	89	116	66
FIFO rp (bps)	2.64	2.54	1.00	0.06	-1.20	-2.49	-3.62	-5.35	-8.63	-18.18	-0.16
LIFO rp (bps)	2.68	2.78	2.68	2.91	2.91	3.12	3.37	3.91	4.85	8.26	2.99
Imbalance (%)	0.8	2.3	3.8	5.5	7.4	9.5	12.0	15.4	20.5	30.2	4.6

Statistical Sensitivity to Order Imbalance: LIFO vs FIFO

In this section, we use a simplified example to show how FIFO can generate statistical bias—that has no economic meaning—in the estimates of realized profitability

compared to LIFO. We examine the case where the LP's inventory is built up over the first n trades of the day followed by D reversing trades—there are a total of D round-trip trades during the day. For simplicity we assume each trade is either an aggressive sale or purchase for 1 share. The LP begins with zero inventory and the initial price of the security is P_0.

To illustrate the statistical bias, we shut down economic sources that can potentially cause differences in the estimates of LIFO and FIFO by assuming that each trade has same price impact Δ in the direction of the trade ($+\Delta$ for buyer-initiated trades and $-\Delta$ for seller-initiated trades), and the occurrence of order imbalance does not convey information about future trades. The first n trades are seller-initiated trades for the security meaning that the LP builds up a cumulative inventory position of $+n$ at time $t = n$ with the security price falling to $P_n = P_0 - n\Delta$. Following the build up, all D subsequent trades are assumed to be aggressive purchases which progressively reverse the LP inventory. In the case when $D = n$ the LP's inventory is completely turned around and there's no trade-imbalance; when $D < n$ the LP ends the day holding $n - D$ shares in inventory.

Figure B.3: LIFO/FIFO Term-Structure Sensitivity

t	Aggressive Sellers	LP	Aggressive Buyers	LP Inventory	ΔPrice	Price
1	Sell →	Buy		+1	- Δ	$P_1 = P_0 - \Delta$
2	Sell →	Buy		+2	- Δ	$P_2 = P_0 - 2\Delta$
3	Sell →	Buy		+3	- Δ	$P_3 = P_0 - 3\Delta$
⋮	⋮	⋮		⋮	⋮	⋮
n-1	Sell →	Buy		+(n-1)	- Δ	$P_{n-1} = P_0 - (n-1)\Delta$
n	Sell →	Buy		+n	- Δ	$P_n = P_0 - n\Delta$
n+1		Sell →	Buy	+(n-1)	+ Δ	$P_{n+1} = P_0 - (n-1)\Delta$
n+2		Sell →	Buy	+(n-2)	+ Δ	$P_{n+2} = P_0 - (n-2)\Delta$
⋮		⋮	⋮	⋮	⋮	⋮
n+(D-2)		Sell →	Buy	n-D+2	+ Δ	$P_{n+(D-2)} = P_0 - (n-D+2)\Delta$
n+(D-1)		Sell →	Buy	n-D+1	+ Δ	$P_{n+(D-1)} = P_0 - (n-D+1)\Delta$
n+D		Sell →	Buy	n-D	+ Δ	$P_{n+D} = P_0 - (n-D)\Delta$

If one were to use FIFO to track the round-trip trades, the presence of an order-imbalance, ceteris paribus, would mechanically generate a downward slopping term-structure. Under FIFO, the trade resulting in the first decrease in inventory at time

$t = n + 1$, is matched with the trade which first increased the inventory position at time $t = 1$. According to FIFO, the LP entered into the position by buying the share at P_1 and later sold it at P_{n+1} to yield a realized profitability of $P_{t+1} - P_1$. More generally FIFO will match the exit trade at time $t = n + d$ with the entering trade at time $t = d$ with realized profitability given by $P_{n+d} - P_d$; note that every round-trip has the same turn-around time of n under FIFO. The average proceeds for a trading day with D round trip trades can be calculated as:

$$\frac{1}{D}\sum_{d=1}^{D}(P_{n+d} - P_d) = \Delta(D + 1 - n) \tag{B.1}$$

Whenever there's trade imbalance, the average turn around time would be decreasing in the average FIFO turn around time n; for example letting $D = \lambda n, \quad \lambda \in (0, 1)$:

$$\partial_n[\Delta(D + 1 - n)] = -(1 - \lambda) < 0$$

So long as there are trade-imbalance days, the FIFO system would have a mechanically downward-sloping term-structure. For days with no-trade imbalance $D = n$, the average turn around time would be n with average proceeds of Δ, regardless of what n is.

Unlike FIFO, LIFO does not have any variation in the realized profitability term structure mechanically introduced by trade imbalance. Since every trade is of the same size, and the inventory reversal begins at time $t = n + 1$, LIFO would match the entering trade at $t = n - d$ with the exit at $t = n + 1 + d$ for $d = 0, 1, 2, \ldots, D$. The round trip times τ for the D trips under LIFO are given by $1, 3, 5, ..., (2D - 1)$. When $D = n$ (no imbalance) the average τ would be the same as FIFO, $\tau_{LIFO} = n = \frac{1}{D}\sum_{d=1}^{D}(2d - 1)$. On days where D is much smaller than n (so imbalance is large) the FIFO τ would be much larger than the LIFO τ, $\tau_{LIFO} = \frac{1}{D}\sum_{d=1}^{D}(2d - 1) << n = \tau_{FIFO}$. Given that each 1-share trade has a price impact of $\pm\Delta$, the price at the entrance and exit may be computed as:

$$P_{n-d} = (P_0 - n\Delta - d\Delta) \quad \text{and} \quad P_{n+1+d} = (P_0 - n\Delta + d\Delta) + \Delta$$

Meaning that the round trip profits, $P_{n-d} - P_{n+1+d} = \Delta$ is constant for every round-trip, resulting in a flat term-structure. Even when trade imbalance is introduced with $D < n$, the realized profitability for every round trip would still remain constant at D. This is to show that in-contrast to FIFO, the combination of price-impact with trade-imbalance does not mechanically generate a downward (or upward) sloping term-structure under LIFO.

APPENDIX C

CO-AUTHORSHIP STATEMENT

Chapter 2 of this dissertation, titled The Profitability of Liquidity Provision, forms the core of a paper of the same name, co-authored with Lingyan Yang and included in this document with her permission.