Would Order-By-Order Auctions Be Competitive?*

Thomas Ernst†, Chester Spatt‡, and Jian Sun§

November 18, 2022

Abstract

Retail trading flow is segregated from non-retail flow in U.S. equities, consistent with cream-skimming. We model theoretically two methods of executing segregated retail trades: a) broker's routing, whereby brokers route orders to market makers and evaluate aggregate performance, and b) order-by-order auctions, where market makers bid on each individual order, a market structure suggested recently by the SEC leadership. We find that order-by-order auctions improve allocative efficiency among market makers, but a winner's curse problem in the auction can reduce retail trader welfare, particularly at times of limited liquidity. Empirical analysis of Retail Liquidity Programs (RLP) currently offered by exchanges shows that these programs behave similarly to order-by-order auctions in our model.

*For helpful comments and feedback we are thankful to: Svetlana Bryzgalova, Pete Kyle, and Mark Loewenstein, along with numerous anonymous industry participants.

†University of Maryland, Robert H. Smith School of Business: ternst@umd.edu

‡Carnegie Mellon University, Tepper School of Business: cspatt@andrew.cmu.edu

§Singapore Management University, Lee Kong Chian School of Business: jiansun@smu.edu.sg

I. Introduction

Retail order flow in U.S. equities is segregated, with retail brokers routing almost all their retail customer orders directly to market makers. These market makers have the option to privately internalize trades off-exchange, typical of economic cream-skimming models. Retail trades are attractive to market makers, either due to lower adverse selection, as in Battalio and Holden (2001), or due to their trades being less correlated, as in Baldauf, Mollner, and Yueshen (2022). In both cases market makers are willing to give retail traders better prices than the exchange due to greater ability to segregate orders. Recently, the SEC Chair has called for a change in market structure to potentially increase competition among market makers.[1] While previous academic work has explored *whether* retail flow should be segregated and its value, once segregated, the question of *how* retail flow should be executed is comparatively unexplored.

We model and evaluate empirically two distinct methods of executing segregated retail trades: broker's routing and an order-by-order auction. Our broker's routing model closely resembles the current market structure, with retail brokers determining where to route each order. While retail brokers use past market maker performance to inform routing decisions, they do not communicate with the market maker prior to routing each individual order. Our order-by-order auction explores possible implementations of the SEC Chair's suggestion (Gensler (2021)), whereby all retail orders enter an auction. This auction would only be available for retail market orders, but each market maker could bid on each individual order.

We evaluate both models with a focus on inventory cost and competition. We assume that each market maker i observes a private signal y_i, and that inventory cost is affected by both the market maker's private signal and the average signal observed by all market makers. Intuitively, each private signal can be thought of as the inventory position of market maker i, with a market maker's willingness to trade depending both on his private inventory and aggregate liquidity of all market makers. We then identify differences in welfare distribution, inventory-management, and order allocation efficiency that arise under each of the two market structures.

In the broker's routing setting, a broker chooses a market maker before the market maker

[1] In remarks before the SIFMA Annual meeting, SEC Chair Gensler stated "I've also sought recommendations around how to instill greater competition for retail market orders on an order-by-order basis, through auctions. With greater competition, more market participants would have access to these retail market orders." Gensler (2022).

observes their private signal. This delivers a highly competitive outcome–with zero expected market maker profits–because market makers are ex-ante identical before observing private signals. When they bid their cost, Bertrand competition obtains, the broker chooses a market maker at random, and market makers earn zero expected profits. This closely mirrors the current system of order routing in equities, where market makers agree to accept order flow from brokers, but there is no pre-trade communication on individual orders. Brokers route to a market maker, and the market maker must accept the order. Evaluation of trades is done on a daily, weekly, or monthly basis, and market makers compete on average rather than bidding against each other on an individual order.

The broker's routing setting delivers strong competition, but the lack of communication on any specific trade means that a trade may be routed to a broker who has observed high ex-post inventory cost. In the order-by-order model, in contrast, brokers bid after observing their signals. In the auction, the market maker with the lowest-cost signal will always win the auction with the most aggressive bid, leading to an increase in allocative efficiency. The common-value nature of the auction, however, creates a winner's curse problem; whichever market maker wins the auction learns that all other market makers had higher signals of cost. Consequently, market makers bid conservatively in the auction.

The welfare of investors can be lower in the order-by-order setting, particularly at times of limited liquidity. Intuitively, once market makers have observed their cost signals, they are more heterogeneous, and Betrand competition does not emerge. Limited liquidity leads to less pressure from auction competitors, less aggressive bids, and larger profits for market makers. While order-by-order auctions have higher allocative efficiency than broker's routing, order-by-order auctions have less competition than broker's routing.

While order-by-order auctions only exist as an SEC concept, we identify a currently-existing close empirical analogue of Retail Liquidity Programs (RLP). Exchange RLP's allow market makers to provide liquidity to retail traders at will by posting hidden limit orders that are only accessible to retail traders. When there is at least 100 shares of RLP interest, exchanges will disseminate an RLP flag indicating the presence of RLP liquidity, though not revealing the exact size or price of the order. If multiple market makers post in an RLP, the market maker with the most aggressively priced order will have first priority for any incoming retail market order, mirroring the potential

competitiveness and allocative efficiency of an order-by-order auction. Unlike the broker's routing system, where market makers must accept any flow the broker routes to them, posting limit orders in a RLP is entirely voluntary: there may be many market makers posting limit orders, or none at all.

Five exchanges currently offer Retail Liquidity Programs. RLPs have times with high levels of market-maker interest, with at least one-sided interest quoted for 20% of the day in Russell 1000 stocks, and over 50% of the day for our sample of liquid ETFs. The trading volume executed in RLPs, however, is small, averaging less than 0.3% of total trading volume, despite exchange trading fees being substantially reduced for trades in the program.

Volume in RLPs is higher in stocks that are tick constrained, consistent with RLPs being particularly effective in more liquid stocks. Volume in RLPs also declines during periods of higher volatility.Order imbalances during times when the RLP flag is active are much lower than order imbalances during the times when the RLP flag is not active, providing further support for the volatility-sensitive nature of voluntary market-maker participation in RLP programs. Price impacts of trades in RLP programs are more sensitive to volatility; when volatility is 1% higher, exchange sub-penny trades have a price impact 8 basis points higher while off-exchange trades have a price impact of only 3 basis points higher. This is consistent with RLP programs ranking lower in the liquidity pecking order of Menkveld, Yueshen, and Zhu (2017), and market makers internalizing trades off-exchange even during periods of higher volatility, consistent with the broker's routing system having no winner's curse.

When the RLP flag is active, mid-quote trading is more common off-exchange as well as on-exchange. The distribution of sub-penny trades is roughly similar, with most of the shift in volume coming from at-quote trading switching to mid-quote trading. These mid-quote trades could come from either retail trading or non-retail hidden liquidity trades; only sub-penny exchange trades are anonymously identifiable as having a retail participant. Quoted bid-ask spreads tend to be more stable when the retail flag is active, consistent with RLPs supplying liquidity during times of high liquidity. When the retail flag is not active, quoted bid-ask spreads tend to be much wider before and after trades.

The SEC idea of an order-by-order auction seeks to "instill greater competition for retail market orders." Under the current system of broker's routing, each order is sent to a single market maker

with no pre-trade communication, and competition is measured by average execution quality. Switching to an order-by-order auction offers a tempting increase in allocative efficiency, as the market makers with the most optimistic signal always wins an auction. But this comes with a drawback, as the market maker who wins by outbidding all competitors with less optimistic signals suffers the auction winner's curse. Market makers scale back their bids, and obtain increased welfare in the order-by-order system. Retail trader welfare can decrease in the switch to order-by-order trading, particularly for volatile stocks and stocks with few competing market makers.

II. Literature and Contribution

Several prior papers argue retail segmentation is optimal as a market design. Battalio and Holden (2001) argue retail traders have lower adverse selection, while Baldauf et al. (2022) argue retail traders are less correlated. Under both cases, it is optimal to segregate retail flow, but different mechanisms for segregating retail flow are not explored. Motivated by the recent SEC calls for order-by-order competition, our paper provides a theoretical analysis into two possible methods of executing retail trades: the current system of broker's routing, and a hypothetical order-by-order system. We show that the proposed order-by-order system would potentially increase allocative efficiency, but decreases retail investor welfare in less liquid stocks.

Parlour and Rajan (2003) argue segmentation decreases consumer welfare, as it leads to a subsidization of retail limit orders. We do not explore the market vs. limit order trade-off, but instead focus on retail marketable orders. Under the SEC vision, retail marketable orders would primarily interact with market maker and non-market maker orders through an order-by-order system similar to the current Retail Liquidity Programs. We empirically analyze these RLP programs and find that they have low liquidity in small stocks and at volatile times, matching our model prediction of how order-by-order systems would function.

One possible analogue to order-by-order trading exists in the option markets, where a considerable share of volume executes in auctions. Bryzgalova, Pavlova, and Sikorskaya (2022) show that these auctions are correlated with retail trading measures, while Ernst and Spatt (2022) present empirical analysis of specific rules, such as a price-match guarantee and out-sized allocation, which prevent competition in option auctions. Hendershott, Khan, and Riordan (2022) present a model and

empirical evidence that auctions in option markets are imperfectly competitive.

Several recent studies have looked at payment for order flow around segmentation. Hu and Murphy (2022), Jain, Mishra, O'Donoghue, and Zhao (2020), and Schwarz, Barber, Huang, Jorion, and Odean (2022) all explore trade-offs in execution quality between PFOF and non-PFOF business models among brokers. Market makers can offer two possible forms of superior prices: PFOF (payments from market makers to brokers) and price improvement (payments from market makers directly to retail customers). Brokers may or may not pass on the total extent of PFOF revenue back to their customers in the form of lower commissions, as documented in Battalio, Jennings, and Selway (2001). In our welfare analysis, we assume that market makers compete solely on price improvement, akin to PFOF being either zero or entirely passed through to retail traders. Our focus is not on the revenue split of PFOF vs. price improvement, but rather what form of market design delivers overall superior or inferior welfare to final retail investors.

Our empirical analysis focuses on Retail Liquidity Programs (RLP) offered by several exchanges. Jain, Linna, and McInish (2021) provide an overview of the NYSE Retail Liquidity Program in 2015. Five exchanges now operate RLPs, and we analyze current RLP data through the lens of learning about potential execution quality under the SEC's proposed order-by-order auctions. They provide a competitive process for both traditional market makers and institutional traders to competitively enter limit orders which offer potential price improvement to retail trades, but empirically suffer from the same winner's curse problem we identify in our theoretical model.

III. Model

The model consists of only two dates, time 0 and time 1, and there is no discounting. There are three types of market players: a retail investor, a broker, and $N \geq 2$ ex-ante market makers indexed by $i \in \{0, 1, ...N\}$. Our focus is the strategic interactions among market makers, so we abstract away from agency problems between the investor and the broker, and assume that the broker's objective is to maximize the investor's profit (which is equivalent to minimizing the spread in our model).

At time 0, the broker receives a one-unit sell order from the investor, and needs to find a market

maker to fulfill the order. [2] We assume that the retail investors is trading for pure liquidity reasons, so there is no information about asset value contained in the direction of the order. If market maker i fulfills the order, it has to hold the position until time 1 which incurs inventory cost c_i. Let s_i be the half bid-ask spread offered by market maker i, then the expected profit that market maker i receives at time 1 is

$$s_i - c_i.$$

The inventory cost c_i is unknown to market maker i at time 0 but has the following structure

$$c_i = c_0 + c_1 \frac{1}{N} \sum_{j=1}^{N} y_j + c_2 y_i,$$

where c_0, c_1 and c_2 are constant, y_i is the private signal observed by market maker i, and is drawn from uniform distribution $U\left[-\frac{1}{2}, \frac{1}{2}\right]$ at time 0. For simplicity, we assume $y_i \perp y_j$ for any $i \neq j$. Intuitively, y_i represents any individual liquidity shocks that market maker i receives at time 0. Each market maker can only observe its own liquidity shock, which affects its inventory cost if executing the retail order. The cost function consists of three components. The first term c_0 is the unconditional expected inventory cost of executing the order, which is the same for all market makers. The second term

$$c_1 \frac{1}{N} \sum_{j=1}^{N} y_j$$

represents its exposure to the aggregate liquidity shock $\frac{1}{N}\sum_{j=1}^{N} y_j$. When c_1 is higher, the inventory cost of executing the order is more sensitive to the aggregate liquidity shock. The third term

$$c_2 y_i$$

represents its exposure to the individual liquidity shock y_i, and similarly, the coefficient c_2 measures the sensitivity of inventory cost to individual liquidity shock. The coefficients $\{c_0, c_1, c_2\}$ are exogenous in our model, and are determined by order characteristics. For example, a stock which is about to announce earnings may have a very high c_1, with market makers very concerned about aggregate inventory imbalances. In contrast, a tick-constrained stock with low informational

[2] The direction of the order does not matter.

asymmetries may have a very low c_1 value, with market makers not very concerned about aggregate inventories.

A. Order-by-order Auction

First, we consider a hypothetical order-by-order auction mechanism. We model order-by-order auction as a common-value auction. Assume that each market maker i submits the spread s_i privately after observing the realization of signal y_i at time 0, so it has the full flexibility of adjusting its spread strategy depending on its assessment of inventory cost. The broker observes spreads offered by all market makers, and chooses the winner with the lowest bid at the end of time 0. If more than one market makers submit the lowest spread, then the winner is chosen randomly among those who submit the lowest spread. At time 1, all players collect their payoffs. We focus on symmetric equilibria such that all market makers choose the same strategy.

Intuitively, when observing a higher signal realization y_i, the inventory cost c_i tends to be larger for market maker i, and thus it will submit a higher spread s_i. We conjecture (and verify later) that there exists a linear symmetric equilibria where all market makers choose the same strategy $s_i(y) = s(y)$ where

$$s = k_0 + k_1 y.$$

We solve the equilibrium using the standard mechanism design approach. Heuristically, suppose all market makers except market maker i follow the equilibrium strategy. We consider market maker i's expected profit $U(z, y)$ where y is the private signal observed by market maker i, and

$$\tilde{s} = k_0 + k_1 z$$

is the spread that market maker i submits to the broker. In equilibrium, we must have

$$\left.\frac{\partial U(z, y)}{\partial z}\right|_{z=y} = 0.$$

The following proposition summarizes our results.

Proposition 1. *There exists a linear symmetric equilibria in which the spread submitted by market*

maker $i \in \{1, 2, ...N\}$ *is*

$$s_i(y_i) = k_0 + k_1 y_i,$$

where

$$k_0 = c_0 + \frac{c_1}{4N}\left(N - 1 + \frac{2}{N}\right) + \frac{c_2}{2N}$$

and

$$k_1 = \frac{N-1}{N}\left(\frac{c_1}{2}\frac{N+2}{N} + c_2\right).$$

Proof. See Appendix.

First, as we discussed earlier, the equilibrium strategy $s_i(y_i)$ is increasing in y_i, with sensitivity

$$k_1 = \frac{N-1}{N}\left(\frac{c_1}{2}\frac{N+2}{N} + c_2\right).$$

This sensitivity coefficient k_1 is increasing in both c_1 and c_2. When c_1 and c_2 are increasing, any market maker i's evaluation of inventory cost is more sensitive to its private signal y_i. As a result, its spread s_i will also be more sensitive to the private signal y_i. Note that this effect still exists even as the number of market makers converges to infinity. The constant term k_0 is an increasing function of all three constants c_0, c_1 and c_2. Intuitively, k_0 is increasing in c_0, as a higher expected inventory cost forces market makers to bid wider spreads. Furthermore, k_0 is also increasing in both c_1 and c_2. This is because when both c_1 and c_2 are increasing, the variation of inventory cost will be higher among market makers. As a result, market makers are ex-post more different from each other, and they are consequently willing to choose a more aggressive equilibrium strategy. The monotonicity of the equilibrium spread also implies that the winner is always the market maker with lowest signal realization, and thus the lowest inventory cost. Order-by-order auctions therefore achieve the first-best outcome in terms of efficient allocation of the retail order, as the retail order is always matched to the market maker with the lowest private cost signal realization. □

B. Broker's routing

In this section, we consider the market equilibrium under broker's routing. The key difference between broker's routing and order-by-order auctions is the flexibility of adjusting spreads ex-post.

Specifically, under broker's routing, market makers are not able to adjust spread order by order. In practice, brokers and market makers establish long-term relationships. Market-maker performance is evaluated on aggregate but not order-by-order, and market makers do not have a choice in when they want to accept order flow from the broker; when a broker sends, they must full-fill the order either by internalizing the order, or paying take fees to fill the order from the exchange. Focusing on this key difference, we model broker's routing by assuming that each market maker i submits the spread s_i before observing the realization of signal y_i, and they are not able to adjust their offered spreads ex-post.

We still focus on symmetric equilibria in this case. In the broker's routing model, market makers are identical when they submit their spreads, as they have not yet observed their private signals. As a result, Bertrand competition obtains, and all market makers will earn zero profit in equilibrium. Therefore, the unique symmetric equilibrium spread must be

$$s_i = \mathbb{E}\left(c_i\right) = \mathbb{E}\left(c_0 + c_1 \frac{1}{N}\sum_{j=1}^{N} y_j + c_2 y_i\right) = c_0$$

for all $i \in \{1, 2, ...N\}$. In equilibrium, all market makers can receive the order with equal probability.

The zero-profit equilibrium is a result of the price competition. While it can be natural to consider quantity competition in broker's routing (Kyle (1985), Baldauf et al. (2022)), with it and the possibility of market makers earning positive profits, but our goal here is to setup a comparable benchmark for order-by-order competition. As our order-by-order auctions feature price competition, we likewise consider price competition for the broker's routing system.

There are several implications. First, although all market makers earn zero expected profit ex-ante, their ex-post profit can be positive or negative, depending on the realized realized inventory cost. In other words, market makers will lose money on some trades. In contrast, the realized profit in the order-by-order auction must be non-negative for all market makers for all trades. Second, under broker's routing, the order is assigned randomly ex-post, as all market makers submit the same lowest spread, and thus will receive the order with equal probability, independent of their realized inventory cost. The order may or may not be executed by the market maker with lowest inventory cost, which incurs a welfare loss in equilibrium. Finally, the equilibrium spread is independent of c_1 and c_2, which are are determined by order characteristics. Broker's routing

therefore delivers much more stable outcomes. The cost reflects c_0, and not differences with market conditions, which motivate our consideration of order-by-order costs c_1 and c_2.

C. Welfare analysis: order-by-order competition vs. broker's routing

In this section, we compare equilibrium outcomes under order-by-order competition and broker's routing. Before doing so, we first introduce definitions for welfare analysis. In both cases, the retail order is always executed, but the inventory cost and equilibrium spread are different. We denote W_M, W_I and W_{total} as the market makers' expected profit, the investor's expected profit and the total welfare, respectively:

1. The expected total profit of all market makers W_M: the expected equilibrium spread minus the incurred inventory cost;
2. The expected total profit of the investor W_I: the expected negative equilibrium spread;
3. The total welfare W_{total}: the expected incurred inventory cost, which is $W_{total} = W_M + W_I$.

Under order-by-order competition, the market maker with the lowest signal realization executes the order in equilibrium, so the expected total profit of all market makers is

$$W_M^{OBO} = \mathbb{E}\left\{\mathbb{E}\left[k_0 + k_1 r - c_0 - c_1 \frac{1}{N}\sum_{j=1}^{N} y_j - c_2 r | \min_i y_i = r\right]\right\}.$$

The investor's expected profit is

$$W_I^{OBO} = -\mathbb{E}\left[k_0 + k_1 r | \min_i y_i = r\right],$$

and the total welfare is

$$W_{total}^{OBO} = \mathbb{E}\left\{\mathbb{E}\left[-c_0 - c_1 \frac{1}{N}\sum_{j=1}^{N} y_j - c_2 r | \min_i y_i = r\right]\right\}.$$

Total welfare is the sum of the welfare of market makers and investors: $W_{total}^{OBO} = W_M^{OBO} + W_I^{OBO}$.

Based on our equilibrium characterizations, we obtain the following Lemma.

Lemma 1. *Under order-by-order competition, the welfare outcomes are*

$$W_M^{OBO} = \frac{1}{N+1}\left(\frac{c_1}{N} + c_2\right),$$

$$W_I^{OBO} = -\left[c_0 + \frac{1}{N(N+1)}c_1 - \frac{N-3}{2(N+1)}c_2\right],$$

$$W_{total}^{OBO} = -\left(c_0 - \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

Proof. See Appendix. □

Order-by-order competition implements the first best allocation, and the total welfare is

$$W_{total}^{OBO} = -\left(c_0 - \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

Note that W_{total}^{OBO} is decreasing in c_0, as c_0 is the expected inventory cost. W_{total}^{OBO} is increasing in c_2, because c_2 determines the variation of inventory cost among all market makers. When c_2 is higher, the expected lowest inventory cost will be lower, and thus the total welfare is higher.

Under broker's routing, all market makers earn zero expected profit, so

$$W_M^{BR} = 0.$$

The equilibrium spread is always c_0, so

$$W_I^{BR} = -c_0.$$

Then the total welfare is

$$W_{total}^{BR} = W_M^{BR} + W_I^{BR} = -c_0.$$

Having solved for the total welfare in the broker's routing setting and order-by-order auction setting, we next compare welfare between the two systems in the following proposition.

Proposition 2. $W_{total}^{BR} < W_{total}^{OBO}$; $W_M^{BR} < W_M^{OBO}$. $W_I^{BR} < W_I^{OBO}$ *if and only if* $\frac{c_1}{c_2} < \frac{N(N-3)}{2}$.

Proof. See Appendix. □

Proposition (2) is a direct result of our previous welfare analysis. Note that the only welfare loss in this setting is from the inefficient inventory management. Under broker's routing, the expected inventory cost is c_0 because the order is randomly assigned ex-post. Under order-by-order competition, the expected inventory cost is

$$c_0 - \frac{N-1}{N+1}\frac{c_2}{2}$$

which is lower than that under broker's routing. The welfare improvement

$$\frac{N-1}{N+1}\frac{c_2}{2}$$

is increasing in N. Intuitively, the expected lowest inventory cost is decreasing in the number of market makers. The lower the lowest-cost market maker, the higher total welfare is under order-by-order auctions. The welfare improvement is also increasing in c_2. As discussed earlier, this is because when c_2 is higher, market makers inventory costs are less correlated, and thus the lowest cost will be lower.

Market maker profit W_M is higher under order-by-order competition, and the difference is also the expected difference of the equilibrium spreads in these two cases:

$$\Delta s = W_M^{OBO} - W_M^{BR} = \frac{1}{N+1}\left(\frac{c_1}{N} + c_2\right).$$

It's clear that Δs is increasing in c_1 and c_2. This is because when c_1 and c_2 are larger, the private signals that market makers observe become more crucial in their evaluation of the inventory cost. Market makers are, effectively, more different ex-post. The ex-post heterogeneity generates the expected positive profit they earn under order-by-order auctions.

The comparison between W_I^{BR} and W_I^{OBO} depends on the level of $\frac{c_1}{c_2}$, and is a result of the trade-off between more efficient inventory management and higher spreads charged by market makers. The welfare gain from more efficient inventory management only depends on c_2, while the increased spread Δs depends on both c_1 and c_2. As a result, the investors benefit from changing from broker's routing to order-by-order auctions only when c_2 is sufficiently large compared to c_1. When this condition is not met, investors are worse off under order-by-order auctions.

A direct result from Proposition 2 is that switching to order-by-order competition may have heterogeneous impacts on stocks with different inventory cost structures. For example, a large stock may be able to absorb a much larger order imbalance than a small stock, that is, for small stock S and large stock L, the market maker's sensitivity to the aggregate imbalance would satisfy $c_1^L < c_1^S$. The market maker's individual cost sensitivity, c_2 could be the same across the stocks, assuming both stocks have the same β on the market and the market maker's individual sensitivity is determined by aggregate exposure to the market. The ratio, of the aggregate-to-individual costs $\frac{c_1}{c_2}$ will be higher for the smaller stock. Consequently, the smaller stock is more likely breach the threshold $\frac{c_1}{c_2} < \frac{N(N-3)}{2}$, and investors in small stocks could have lower welfare under order-by-order auctions compared to broker's routing.

We do not directly model endogenous entry of market makers, but our model gives implications for how market competition and liquidity provision changes welfare outcomes in partial equilibrium analysis. Proposition 2 implies that when the number of market makers N is small, the investor's welfare likely lower upon switching to order-by-order auctions. Here N measures the number of active market makers who provide liquidity. During time periods when market makers are not willing to provide liquidity (for example, due to market uncertainty or high inventory cost), our model predicts that switching to order-by-order competition will be more likely to hurt investors. For example, if investor protection is more important during market distress, our result highlights the unintended negative effect of order-by-order auctions during time periods when investors face high uncertainty.

IV. Retail Liquidity Programs

Several exchanges have developed retail liquidity programs (RLP). These programs enable market makers to enter hidden limit orders which improve on the NBBO but are only accessible to retail orders. These hidden limit orders can be priced in any one-tenth of a cent increment, with the exception of the IEX RLP which only allows pricing at the mid-quote. While the orders are hidden, if there is more than one round lot of interest at a price which improves the NBBO, the exchange will disseminate an indicative flag highlighting that there is a resting limit order, but it will not indicate the price or size of the order.

Retail Liquidity Programs offer the closest existing analogue to the contemplated order-by-order competition system for retail trades. If multiple market makers place limit orders, for example, the market maker with the best priced limit order will win any incoming retail market order. In approving RLPs, the SEC itself has frequently highlighted the same objectives that it has for proposing order-by-order competition, namely to increase the number of market participants interacting with retail trades. These RLP limit orders are only available to retail traders, preserving the cream-skimming separation of retail traders, but also have entirely voluntary participation: market makers may chose to stop bidding for incoming orders at any time. We analyze RLPs as a way to gain insight into how the order-by-order system would function, and identify similarities between our model and the current utilization of RLP programs.

A. Program Details

NYSE was the first to operate a RLP, on August 1, 2012.[3] The NYSE RLP was initially approved as a pilot and given several temporary pilot extensions until permanent approval on February 15, 2019. Any NYSE member can submit a Retail Price Improvement Order (RPI). An RPI order can be submitted in \$0.001 increments, and must improve the best bid or offer on the NYSE or NYSE Arca book by at least \$0.001. The size and exact price of resting RPI orders are non-displayed, but the orders do trigger indicative messages on the SIP and NYSE proprietary data feeds indicating whether there is any RPI interest at the ask, any RPI interest at the bid, or any RPI interest at both. Incoming marketable retail orders can trade against resting RPI orders. Incoming retail orders will first trade against the best-priced orders; if there is a non-displayed order which is not RPI at the mid-quote, the retail order would trade against the mid-quote interest before trading against any RPI orders priced between the mid-point and near side. Retail marketable orders can be set to only trade against RPI and non-displayed orders, or to trade against any RPI and non-displayed orders and then subsequently against the displayed best quotes up to the limit price.

The NYSE Retail Liquidity Program charges no trading fee to qualifying retail market orders. The NYSE RLP program also pays \$0.0003 credit to a Retail Liquidity Provider whenever their

[3]The introduction of the data field for the RLP led to the \$400 million trading glitch at Knight Capital Group on the first day that the new data field was active.

RPI limit order fills a retail market order. To qualify as a Retail Liquidity Provider on the NYSE, a firm must maintain a resting RPI order which improves the best bid or offer for at least 5% of the trading day.

This 5% rule distinguishes the NYSE Retail Liquidity Program from those offered by NASDAQ and BATS, with both competing programs being developed shortly after the NYSE program. The BATS program was approved as a pilot on November 27, 2012, while the NASDAQ program was approved as a pilot on February 15, 2013. Both programs have no requirement to provide liquidity for a certain percentage of the trading day, it is more accessible to non-market-making firms. In approving the NASDAQ RLP, the SEC notes that "the Program might also create a desirable opportunity for institutional investors to interact with retail order flow that they are not able to reach currently. Today, institutional investors often do not have the chance to interact with marketable retail orders that are executed pursuant to internalization arrangements. Thus, by submitting RPI Orders, institutional investors may be able to reduce their possible adverse selection costs by interacting with retail order flow" SEC (2013). The SEC identifies the same desirable feature, that of more potential counter parties for retail trades, that are highlighted in a potential move to order-by-order competition.

The Investors Exchange (IEX) offers a retail liquidity program whereby retail liquidity providers can enter hidden mid-point peg limit orders which are only available to retail market orders. All mid-point peg orders enter the same time priority queue, whether or not they are only available to retail traders, and both have queue priority over the IEX D-limit order, which is the discretionary limit order which takes advantage of the IEX speed bump to reprice when it detects a crumbling quote. The IEX RLP only takes mid-point orders, and disseminates a RLP indicative flag when there is at least one round lot of RLP interest. All eligible retail orders have no trading fees, either for the retail broker or the retail liquidity provider.

The IEX RLP is the most recent, with SEC approval granted on September 29, 2021. Unlike other retail liquidity programs, the IEX program only allows mid-quote prices. Therefore, while the size available is hidden, an advertised RLP indicator from IEX confirms that at least 100 shares are available at the specific price of the mid-quote. As a result, the program required an approved exemption from SEC Rule 242.602, as the RPI would indicate a specific price and a minimum quantity of shares, but would not be accessible to non-retail marketable orders.

The Members Exchange MEMX applied to create an RLP program but was denied by the SEC on February 14, 2022. The MEMX proposal differed from previous proposals in the determination of price-time priority. Under the MEMX proposal, incoming retail market orders first interact with hidden RPI orders before interacting with hidden non-RPI orders, even if the hidden non-RPI are at the same price level and have time priority. MEMX argued that because hidden PRI orders do contribute to the dissemination of the RPI interest indicator, they should have priority over hidden non-RPI orders at each price level, analogous to standard practice of non-hidden orders having priority over hidden orders at each price level. The SEC disagreed, and ruled that the change in prioirty would violate Section (6)(b)(5) and Section 11A of the Exchange Act.

B. Data and Summary Statistics

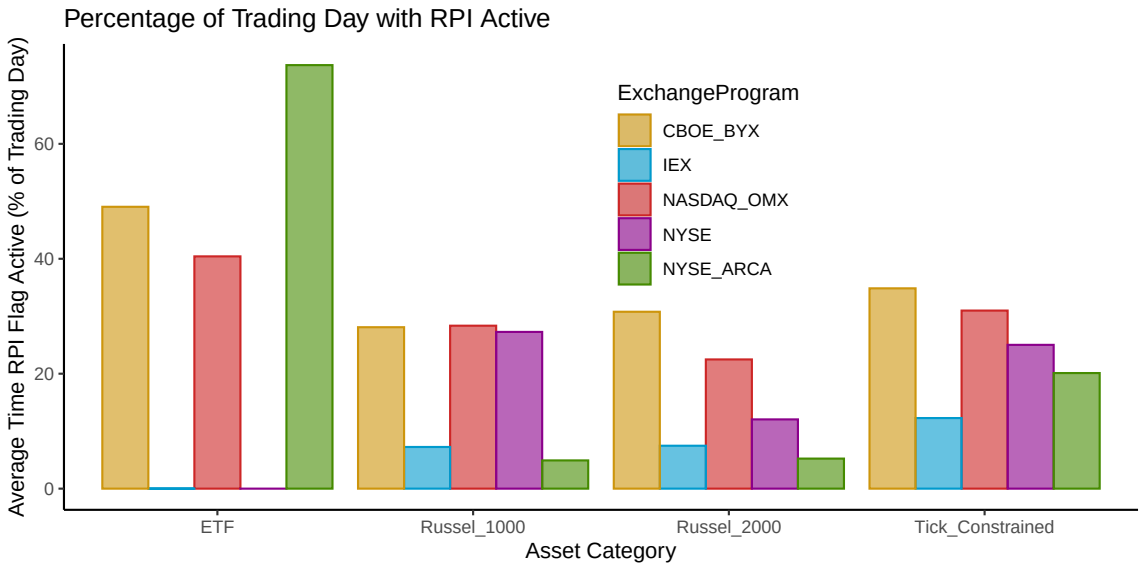
We obtain NYSE TAQ (Trade and Quote) data from January 1, 2022 to May 30, 2022. We examine all securities in the Russell 3000 index, as well as the 100 most frequently traded ETFs, provided these securities are priced above $1 per share. To exclude fractional shares, as documented by Bartlett, McCrary, and O'Hara (2022), we exclude any orders for exactly 1 share, as these may be orders for a fractional share which rounds up to 1 share.

Retail Liquidity Program (RPL) Indicators are distributed through the SIP, and are available in TAQ Data. As indicators may be disseminated even when an exchange's visible posted best bid or offer (BBO) is not at the official NBBO, we obtain retail indicator flags from the TAQ Quotes file. For each trade occurring at an exchange with an RLP, check whether the trade occurred with an active RLP flag by matching the RLP quotes for that exchange using the participant timestamp. We also construct an indicator for whether any RLP from the five different programs is active at any point in time, and match this to both on-exchange and off-exchange trades using the participant timestamps.

Retail Liquidity Programs have indicative interest for a large portion of the trading day. Figure 1 plots the percentage of time, by asset, that there is at least one-sided RPI interest. ETFs often have resting RPI orders for 50 to 75% of the trading day. For stocks of the Russell 1000, NYSE, CBOE, and NASDAQ have resting RPI orders for over 20% of the day. For stocks of the Russell 2000, both CBOE and NASDAQ have resting RPI orders for over 20% of the day. While some of the differences in RPI shares across assets may come from the different rules, as outlined in section

V.A, there is also a considerable listing-exchange advantage. NYSE Arca's retail liquidity program, for example, has RPI interest for less than 10% of the trading day for Russell 1000 or Russell 2000 stocks, but has RPI interest for over 75% of the trading day for ETFs in our sample.

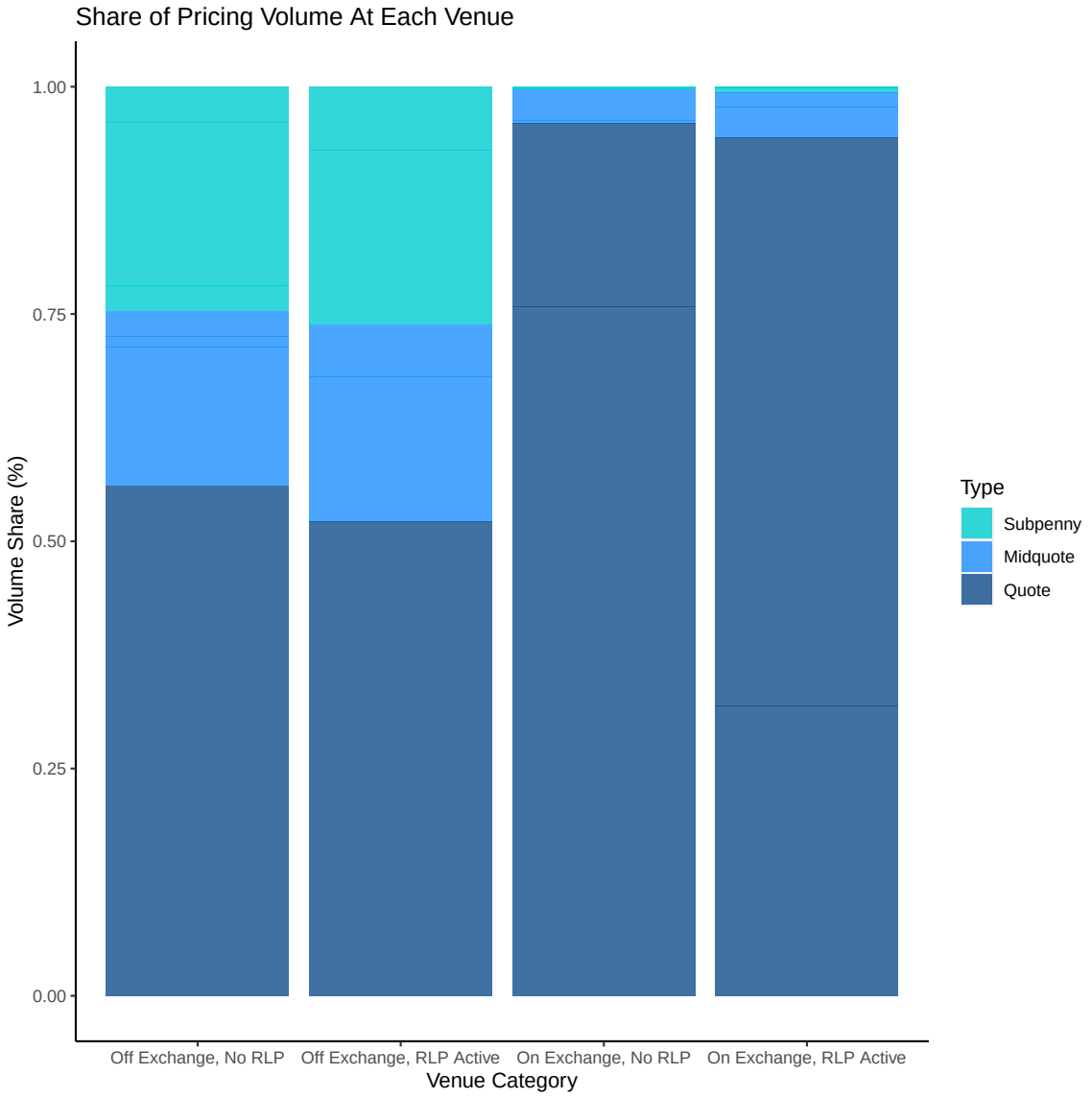
Figure 1. Time Share of Retail Liquidity Programs. We plot the average time that an RPI indicator is active, measured as percentage of time active out of the total trading day. Our sample can be divided into three groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the the day for at least one-third of the days of our sample.



While the percentage of the trading day with RPI interest is considerable, the volumes executed through RPI programs are diminutive. Figure 2 depicts the trading volume split of trades when the exchange's RLP flag is, and the trading volume split when it is not. Sub-penny executions are less than 1% of total trading volume at exchanges, even when the RLP flag is active. On-exchange mid-quote trading volume is considerably higher when the RLP flag is active, but still represents less than 5% of total trading volume. This mid-quote volume is a mixture of both retail interest, including from the IEX RLP which only allows retail RPI orders to be priced at mid-quote, and non-RLP program hidden mid-quote liquidity. The vast majority of mid-quote and sub-penny trading occurs off-exchange. When RLP flags are active, a larger share of off-exchange volume occurs at sub-penny or mid-quote prices. Table I presents the exact total volumes in our sample executed when RPI programs are active, and when they are not. Note that exchange sub-penny volume when there is no RLP Flag is small but non-zero, suggesting imperfect matching between the

RLP Flags disseminated through the SIP and the exchange trades, as matched on their participant timestamps. One possible explanation for this discrepancy is given by Schwenk-Nebbe (2021), who show that the exchange processing and dissemination of quotes is typically several microseconds faster than that of trades.

Figure 2. Volume Share of Venues. We plot the percentage of volume which executes either at the quote, at the mid-quote, or at a sub-penny price for both on-exchange and off-exchange venues. On both types of venues, a higher percentage of volume occurs at the quote when there is no RLP flag active, and a higher share of volume executes at sub-penny and mid-quote prices when the RLP flag is active.



The total volume share of Retail Liquidity Programs is stable during our sample period. As

Panel A of Figure 3 depicts, on-exchange sub-penny retail trades are consistently less than 0.1% of total volume for the Russell 1000 and Russell 2000 stocks in our sample. The volume share of ETFs and tick-constrained stocks is slightly higher, at around 0.2% to 0.3% of total trading volume. We define a stock as tick-constrained if it has a one penny bid-ask spread for at least 50% of the trading day for at least one-third of the trading days in our sample. For these stocks, competition for a marketable order is potentially larger due to the tick constraint, with increased interest in providing liquidity in an RLP. In Panel B of Figure 3, we plot the volume of any exchange sub-penny or mid-quote executions while the RLP flag is active. While this will include some non-retail hidden liquidity, it also captures retail interest at mid-quote, which is crucial as the IEX RLP only allows pricing retail price improvement at mid-quote. For ETFs and tick-constrained stocks, exchange sub-penny and mid-quote volume when the RLP flag is active is around 0.5% to 0.9% of trading volume. While this is a small share of total trading volume, it represents a much larger fraction of retail trades.

Table I: Summary Volumes At Each Price Increment. This table presents summary total trading volume (in billions of dollars) in our sample for each sub-penny category of trade: at-quote, mid-quote, and sub-penny. Panel A of Figure presents volume for exchange trades. We define an RLP Flag as active if there is contemporaneous RLP interest at the exchange where trade occurs. Panel B presents the volume for off-exchange trades. We define an RLP Flag as active if there is contemporaneous RLP interest at any exchange with an RLP program.
Our sample can be divided into three asset groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) asset group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the the day for at least one-third of the days of our sample.

Panel A: Exchange Trades

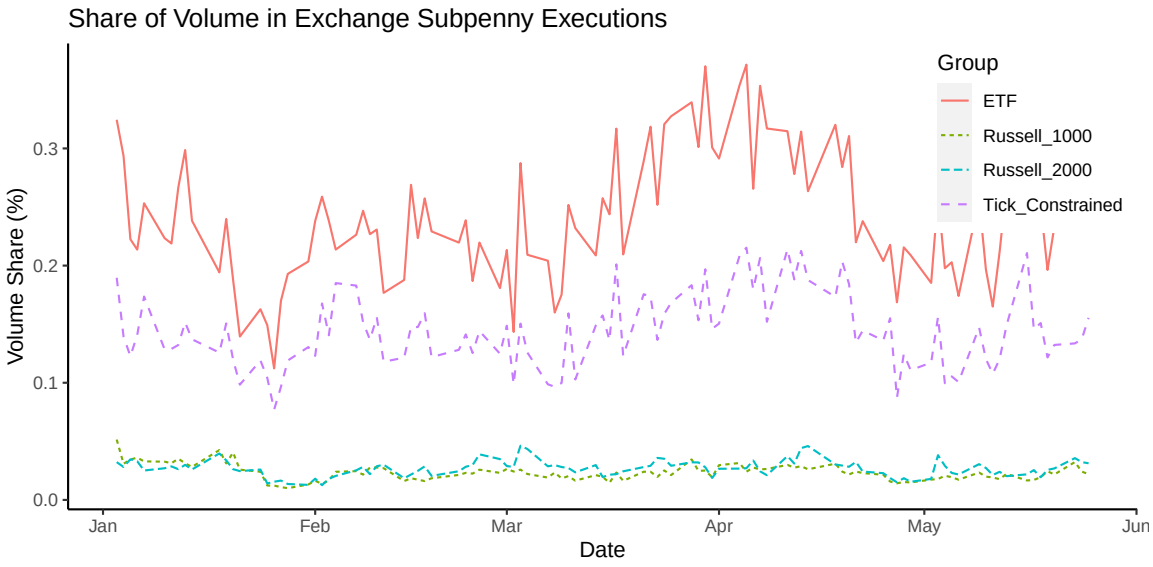
Asset Class	RLP Flag	Mid-quote Volume	At Quote Volume	Sub-penny Volume
ETF	Active	367.55	7511.59	84.07
ETF	None	50.93	1974.41	10.39
Russell_1000	Active	427.25	8721.88	31.71
Russell_1000	None	341.05	8989.47	22.34
Russell_2000	Active	43.45	761.53	0.87
Russell_2000	None	43.98	813.33	0.18
TickConstrained	Active	332.61	5208.34	35.94
TickConstrained	None	106.13	2097.01	11.51

Panel B: Off-Exchange Trades

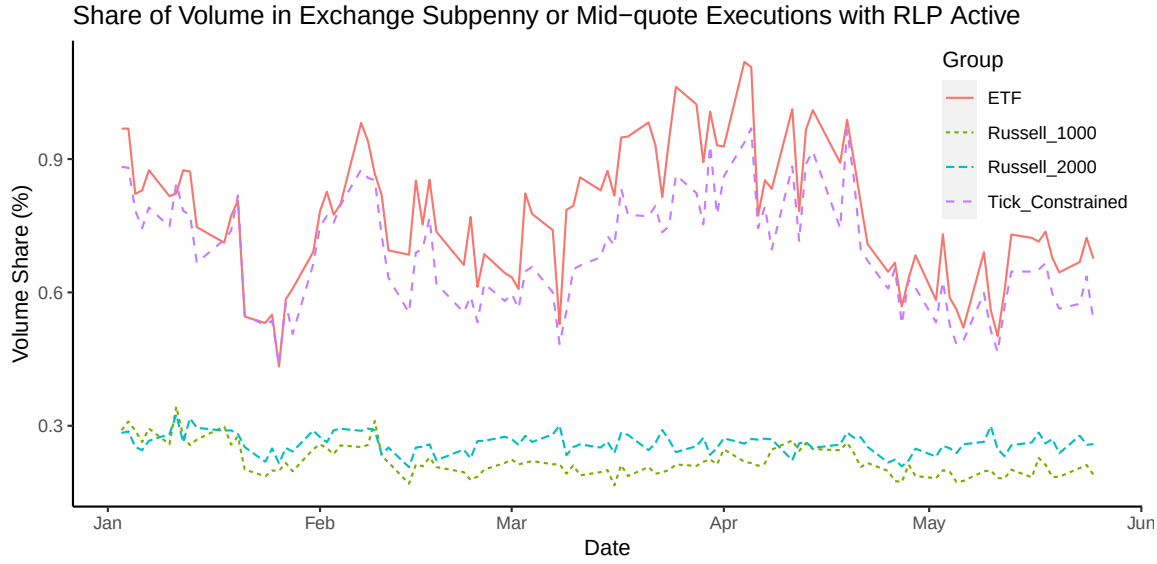
Asset Class	RLP Flag	Mid-quote Volume	At Quote Volume	Sub-penny Volume
ETF	Active	676.05	1751.87	1023.03
ETF	None	114.15	385.90	236.14
Russell_1000	Active	1242.00	3111.26	1279.47
Russell_1000	None	1186.18	3571.36	1447.22
Russell_2000	Active	98.70	262.64	82.13
Russell_2000	None	100.20	287.40	82.74
TickConstrained	Active	594.96	1150.28	759.10
TickConstrained	None	240.86	539.94	334.39

Figure 3. Volume Share of Sub-penny Retail Liquidity Programs. For each day, we plot the share of volume which executes in a retail liquidity program, out of total volume. Our sample can be divided into three groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the the day for at least one-third of the days of our sample. Panel A presents the volume share of only exchange sub-penny executions while a RLP indicator is active, while Panel B presents the volume share of all exchange sub-penny or mid-quote executions while a RLP indicator is active.

Panel A: Volume Share for Exchange Sub-penny Executions with an RLP Flag Active



Panel B: Volume Share for Exchange Sub-penny or Mid-quote Executions with an RLP Flag Active



C. *RLP Programs and Adverse Selection*

Under the broker's routing model, market makers must accept order flow from brokers. Market makers do have a choice in where to execute the trade, either by internalizing the trade, or sourcing liquidity from an external venue like an exchange or dark pool. The choice to externalize, however, is not without cost: market makers will fail to capture the spread on any trades they externalize, must pay PFOF if the broker charges PFOF, and will have to pay any trading fees associated with trading on an external venue. For marketable orders, these fees are generally positive. The order-by-order model, in contrast, is entirely at will. Market makers can post liquidity when they desire to do so, and may withdraw their quotes when they do not.

The current structure of exchange retail liquidity programs has this same at-will feature of liquidity provision, with liquidity-providing participants in the program under no obligation to guarantee execution of retail trades.[4] As a result, exchange retail programs offer insight into the potential workings of an order-by-order model, where market makers are under no obligation to participate for all orders. While many market makers may wish to provide liquidity for orders in large stocks during periods of low volatility, our model suggests that periods of high volatility or in small stocks. Motivated by this reasoning, we estimate the following regression.

REGRESSION 1: *For each asset i on date t with retail flag f:*

$$\begin{aligned} VolumeShare_{ift} = \quad & \alpha_0 + \alpha_1 Percent_Time_At_Minimum_Spread_{it} + \alpha_2 Volatility_{it} \\ & + \alpha_3 Average_Volumei + \alpha_4 Absolute_Intraday_Return_{it} + \epsilon_{ijkt} \end{aligned}$$

Results of Regression 1 are presented in Table II. Sub-penny trading volume, as a percentage of all trading volume, is considerably larger when assets spend a larger percentage of the day at the minimum bid-ask spread. Stocks with larger average trading volume also consistently have larger sub-penny volume. Exchange sub-penny volume, however, is decreasing in stock volatility. Days with higher volatility have less on-exchange sub-penny trading, and more off-exchange sub-penny trading. Under they hypothesis of Boehmer, Jones, Zhang, and Zhang (2021) that sub-penny trades

[4]We note that the NYSE RLP does have a requirement that retail liquidity providers provide price-improving RPI limit orders for at least 5% of the trading day on a certain fraction of trading days to qualify for superior trading fee / rebate pricing. As Figure 1 makes clear, this threshold is low compared to the percentage of time that RPI orders are active.

capture retail trading, retail trades tend to execute off-exchange rather than on-exchange in times of higher volatility, suggesting a direct commitment by market makers to provide liquidity when it is needed.

Table II: Cross-Sectional Variation in Volume Shares. This table estimates Regression 1 with sub-penny volume, measured as a percentage of total daily trading volume. We divide sub-penny volume into three groups: on-exchange (which must have an active retail flag), off-exchange under an active retail trade, and off-exchange with no retail flag. For stock i on date t, $Percent_Time_At_Minimum_Spread_{it}$ measures the percentage of the trading day with a quoted bid-ask spread of one penny, $Volatility$ measures the standard deviation of 15-minute returns, $Average_Volume_i$ measures the average daily trading volume in the security over our entire sample period, and $Absolute_Intraday_Return_{it}$ measures the absolute value of the intra-day return.

	Dependent variable: Sub-penny Volume		
	Exchange Retail Flag	Off-Exchange Retail Flag	Off-Exchange No Retail Flag
	(1)	(2)	(3)
Time At One Tick	0.001***	0.039***	−0.005***
	(0.0001)	(0.003)	(0.002)
Volatility	−0.022***	0.175	0.999***
	(0.006)	(0.156)	(0.102)
Average Volume	0.032***	0.247***	0.149*
	(0.012)	(0.075)	(0.080)
Absolute Intra-day Return	−0.003	4.624***	4.847***
	(0.045)	(1.038)	(0.887)
Observations	259,081	259,081	259,081
R^2	0.012	0.101	0.031
Adjusted R^2	0.012	0.101	0.031
Residual Std. Error (df = 259076)	0.435	2.988	2.433
Note:			*p<0.1; **p<0.05; ***p<0.01

While the Retail Liquidity Programs are the only way that on-exchange trades can be priced in sub-penny increments, retail trades can trade in a variety of methods, with Barber, Huang, Jorion, Odean, and Schwarz (2022) estimating that less 35% of retail trading takes place at sub-penny prices. We directly compare on-exchange sub-penny trades with off-exchange sub-penny trades, as these off-exchange trades are the closest analogue to on-exchange trades. Barardehi, Bernhardt, Da, and Warachka (2022) document, however, that sub-penny trading may be driven

not by the activity of retail traders, but instead indirectly reflect price improvement given when retail traders are willing to take the opposite position of institutional flow. Under this hypothesis, the increased off-exchange sub-penny pricing suggests market makers increase their role in intermediating between retail traders and institutional flow. Conversely, the level of on-exchange sub-penny pricing decreases, which is consistent with institutional traders or market makers reluctance to offer liquidity to any retail trades, consistent with the winner's curse we identify in our model.

Figure 4 depicts the distribution of order imbalances when the RLP flag is or is not active. When the retail flag is active, order imbalances are tightly clustered around a near-zero imbalance, with as many buys arriving as sells. When the retail flag is not active, order imbalances have a distribution with a much larger variance, with a much greater likelihood of large positive or negative order imbalances. This is consistent with the entirely discretionary nature of the RLPs. The SEC views the opportunity for "institutional investors to interact with retail flow" as desirable[5], but it is important to note that institutional traders may be eager to buy from retail traders at times, or sell to retail traders at times, but unlikely to want to stand ready to buy or sell to retail traders at any time on demand.

Retail traders placing market orders may arrive at any time in the day, including during periods of stress. Even under the most generous assumption that their orders are uncorrelated with aggregate institutional order flow, half their order flow would be in the same direction as aggregate institutional order flow. To investigate the relationship between retail liquidity program volume and price movements, we re-estimate Regression 1 with price impacts, and present the results in Table III. In periods of higher volatility, all sub-penny trades have increases in price impact, but exchange sub-penny trades have the largest increases in price impact. This is consistent with the decrease in volume of exchange trades, and the only RPI interest in retail liquidity programs being from interest in trading against one side of the retail flow, and not two-sided liquidity provision.

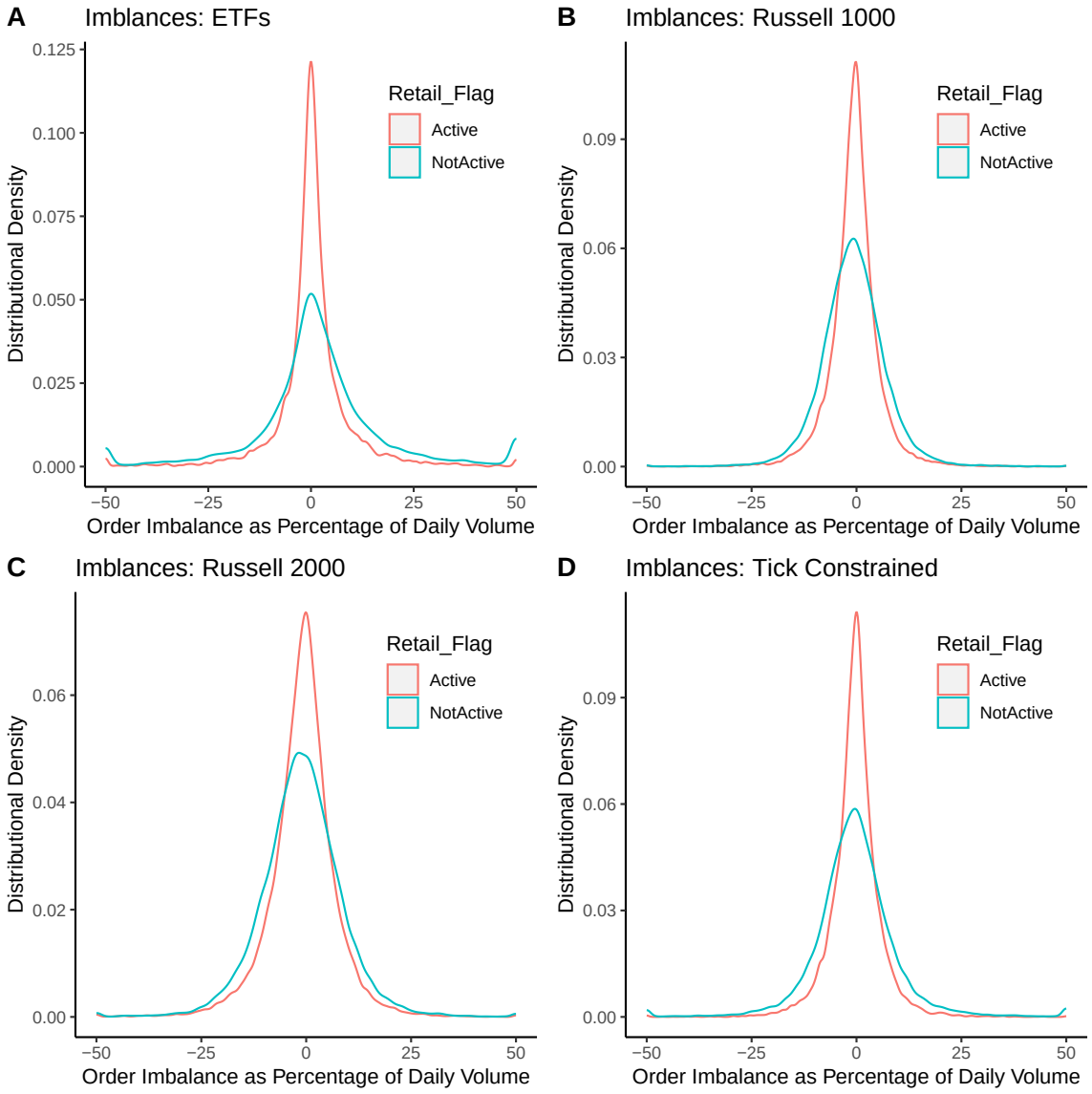
Along with the volume of liquidity provided, there is also the consideration of the prices given. We plot the distribution of sub-penny prices for a single trading day in Figure 5. For both on-exchange and off-exchange trades, there is more mid-quote volume when the RPI flag is active compared to when there is no active RPI flag. For off-exchange trades, however, the distribution of sub-penny price improvement is strikingly consistent between the two times. At each possible level

[5]See SEC (2013).

Table III: Cross-Sectional Variation in Price Impact. This table estimates Regression 1 with mean price impact of sub-penny trades, measured in basis points at 30 seconds following a trade. We divide sub-penny trades into three groups: on-exchange (which must have an active retail flag), off-exchange under an active retail trade, and off-exchange with no retail flag. For stock i on date t, $Percent_Time_At_Minimum_Spread_{it}$ measures the percentage of the trading day with a quoted bid-ask spread of one penny, $Volatility$ measures the standard deviation of 15-minute returns, $Average_Volume_i$ measures the average daily trading volume in the security over our entire sample period, and $Absolute_Intraday_Return_{it}$ measures the absolute value of the intra-day return.

	Dependent variable: Sub-penny Price Impact (BPS)		
	Exchange Retail Flag	Off-Exchange Retail Flag	Off-Exchange No Retail Flag
	(1)	(2)	(3)
Time At One Tick	-0.018^{***}	-0.005^{*}	-0.053
	(0.003)	(0.003)	(0.050)
Volatility	7.955^{***}	2.960^{***}	3.512^{***}
	(0.851)	(0.213)	(0.769)
Average Volume	-0.201^{**}	-0.103^{**}	-0.034
	(0.086)	(0.043)	(0.068)
Absolute Intra-day Return	-5.527	-4.509^{***}	-2.717
	(8.850)	(1.602)	(3.688)
Constant	0.668^{**}	0.278^{**}	0.207^{**}
	(0.303)	(0.119)	(0.096)
Observations	61,960	248,481	255,384
R^2	0.046	0.0002	0.0002
Adjusted R^2	0.046	0.0002	0.0002
Residual Std. Error	14.148 (df = 61955)	71.025 (df = 248476)	119.257 (df = 255379)
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 4. Distribution of Order Imbalances. For each stock-day observation, we calculate the total order imbalance among trades occurring when the RLP flag is active, and the total order imbalance among trades occurring when the RLP flag is not active. For stock i on date t with flag j, imbalance is calculated as $Imbalance_{ijt} = \frac{\sum Buy_{ijt} - \sum Sell_{ijt}}{\sum Buy_{ijt} + \sum Sell_{ijt}}$. We plot the distribution of imbalances, with the tails truncated to an imbalance of $\pm 50\%$. Panel A presents the imbalance distribution for ETFs, Panel B for stocks in the Russell 1000 Index, Panel C for stocks of the Russell 2000 Index, and Panel D for stocks and ETFs which are tick-constrained, defined as having at a one-penny bid-ask spread least 50% of the trading day for at least one-third of the trading days in our sample.

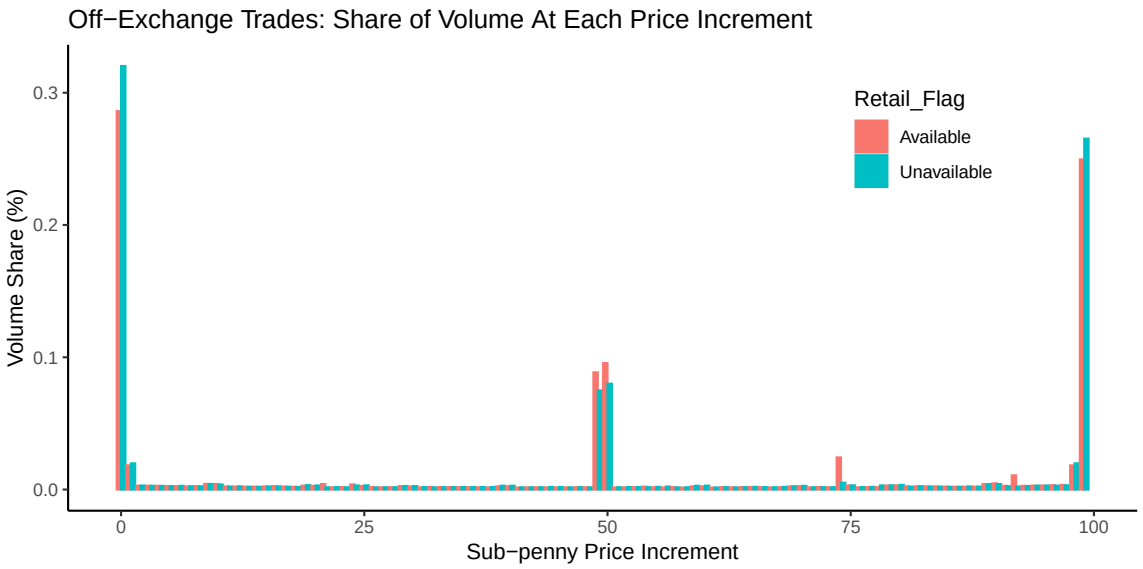


of sub-penny pricing, for trades receiving between 2 and 48 mils of improvement, and for trades between 52 and 99 mils of improvement, the share of off-exchange volume is the same whether or not the RPI flag is active, consistent with market makers delivering a consistent fill to retail market

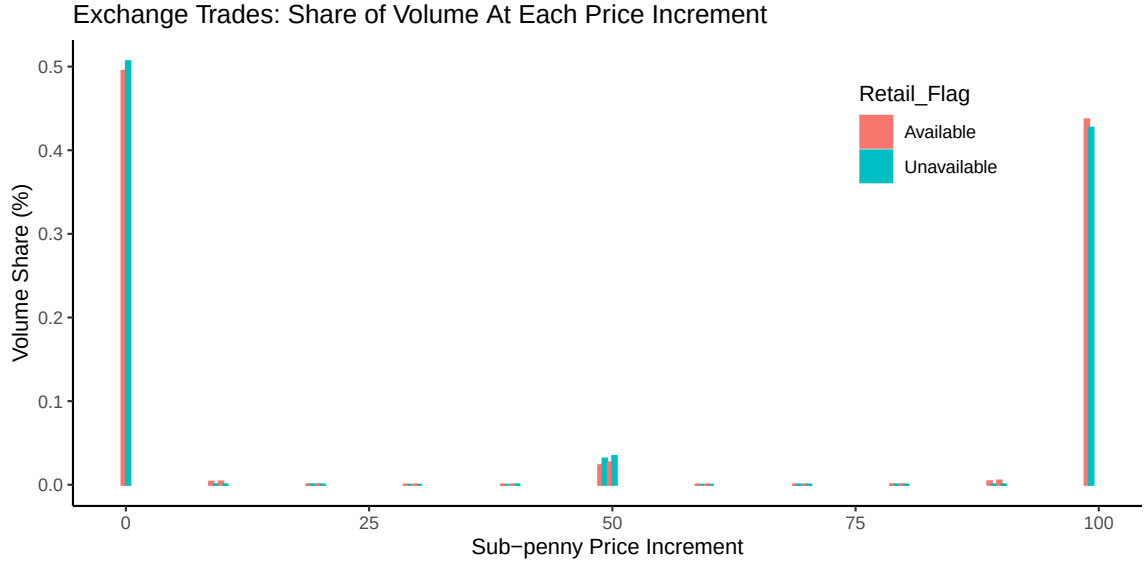
orders.

Figure 5. Volume Share of Off-Exchange Sub-Penny Prices. For each possible sub-penny price increment, we plot the volume occurring at this price increment, as a percentage share of all volume occurring on that venue. We use data from trades occurring on January 3, 2022, and separate trade volume into two categories: trades occurring when there is a retail liquidity program flag active, and trades occurring when there is no retail liquidity program active. Panel A presents the distribution of trades for off-exchange trades. Panel B presents the distribution of trades for on-exchange trades. Note that on-exchange sub-penny trades can only occur in increments of tenths of a cent.

Panel A: Price Improvement Distribution for Off-Exchange Trades



Panel B: Price Improvement Distribution for On-Exchange Trades

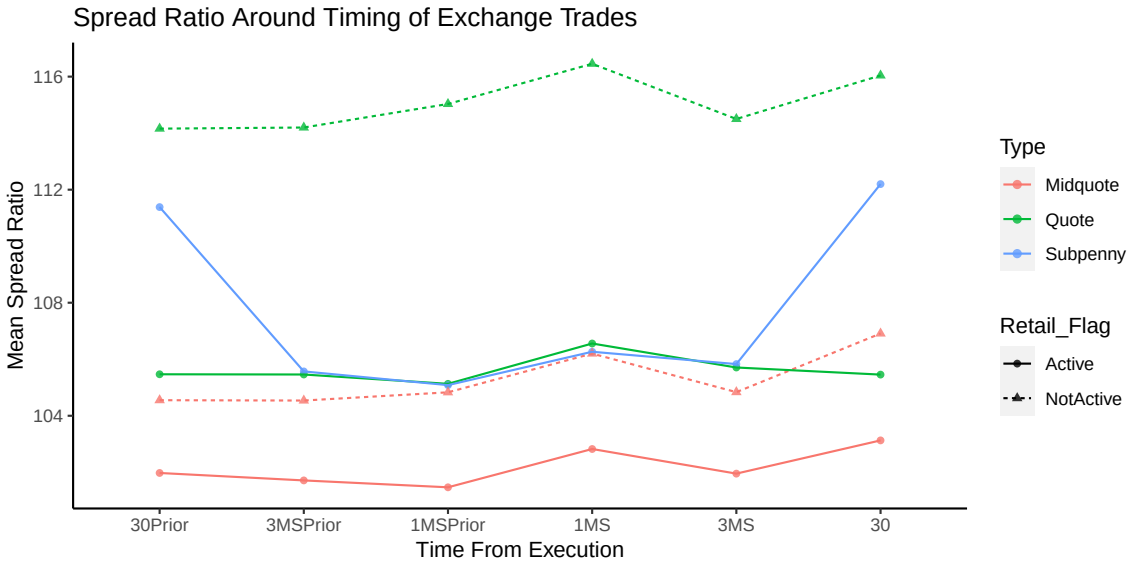


We also investigate the interaction between trading volume and quoted bid-ask spreads, both when the RPI flag is active and inactive. Figure 6 presents the ratio of quoted spreads before an after trades. We first divide trading volume into on-exchange and off-exchange trades, and then further divide volume into sub-penny, mid-quote, and at-quote bins. For each individual stock, we observe the quoted bid-ask spread q_{t+i}, where i can be ± 30 seconds, ± 3 milliseconds, or ± 1 milliseconds. We then calculate the average spread $\bar{q}_{t+i}$ separately for when the RPI flag is or is not active, and plot the ratio $\frac{\bar{q}_i}{\bar{q}}$. When the retail flag is active, off-exchange spreads are very stable, with the same bid-ask spread before and after a trade. When the retail flag is not active, off-exchange spreads before and after a trade tend to be around 2 to 4% wider on average, for all categories of pricing.

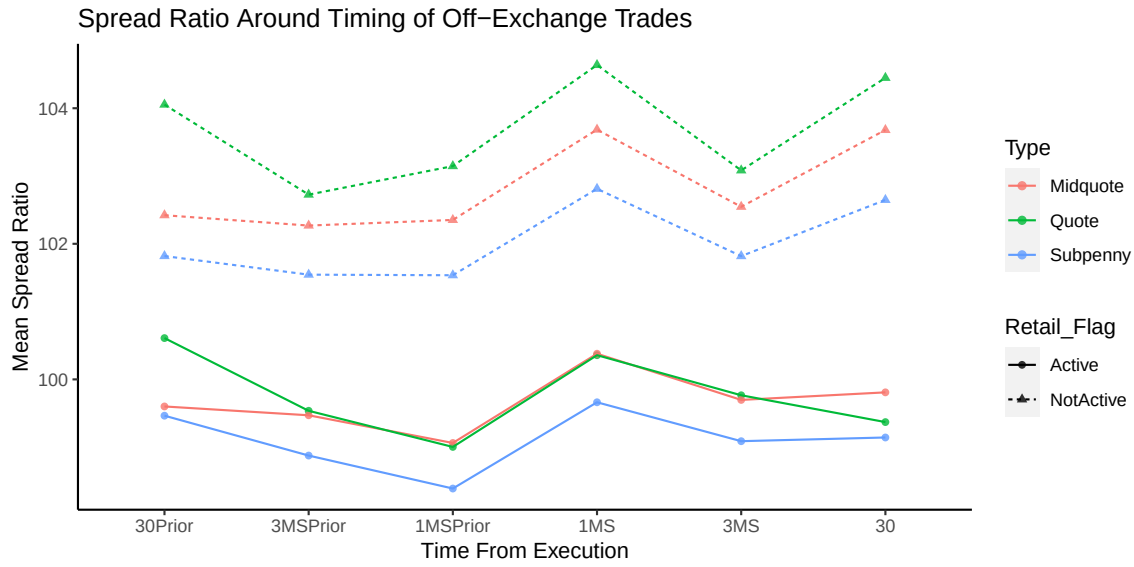
Exchange trades have wider spreads overall before and after trade, consistent with the liquidity-demanding nature of at-quote trading, and consistent with the pecking order proposed by Menkveld et al. (2017). The exchange sub-penny trades when the retail trade is active have a spread ratio even higher than that for exchange trading, suggesting that exchange retail liquidity programs generally rank below off-exchange trading in a pecking order of venues, even when the retail flag on-exchange is active.

Figure 6. Spread Ratios Around Trades. We plot the change in spreads around the timing of a trade, separately for trades occurring when the RLP flag is active, and trades occuring when it is not active. For trades occurring at time t, we calculate the quoted spread q_t as well as the quoted spread q_{t+i} occurring at a time-offset of i. We then calculate the mean quoted spreads $\bar{q}$ and $\bar{q}_{+i}$, and plot their ratio $r = \frac{\bar{q}_i}{\bar{q}}$. We consider time offsets of 30 seconds prior to trade, 3 milliseconds prior to trade, 1 millisecond prior to trade, 1 millisecond after trade, 3 milliseconds after trade, and 30 seconds after trade.

Panel A: Spread Ratio For Exchange Trades



Panel B: Spread Ratio For Off-Exchange Trades



V. Conclusion

In the current market structure, retail brokers set up relationships with market makers, and send individual orders to individual market makers. While market makers are evaluated on the average execution quality they deliver, there is no communication over individual orders. The SEC concept for order-by-order auctions would instead have market makers bid on individual orders.

Our model shows that a switch to order-by-order auctions comes with trade-offs. Allocative efficiency is improved, as order-by-order auctions ensure that an incoming retail market order is always routed to the market maker who has observed the lowest cost signal. Given the common-value nature of the auction, however, there is a winner's curse. Market makers obtain higher profits in the auction relative to the broker's routing system. Retail traders can be worse off in the switch to order-by-order auctions, particularly in times of high volatility or in illiquid stocks.

We empirically evaluate Retail Liquidity Programs (RLPs) to gain insight into how an order-by-order auction would function. While these RLPs offer potential price improvement to retail orders, this liquidity disappears in times of volatility. As in our theoretical model of order-by-order auctions, observed trades in RLP programs tend to occur at times of lower volatility, and times when order imbalances are smaller.

REFERENCES

Baldauf, Markus, Joshua Mollner, and Bart Z. Yueshen, 2022, Siphoned Apart: A Portfolio Perspective on Order Flow Fragmentation, *Available at SSRN 4173362* .

Barardehi, Yashar, Dan Bernhardt, Zhi Da, and Mitch Warachka, 2022, Institutional Liquidity Demand and the Internalization of Retail Order Flow: The Tail Does Not Wag the Dog .

Barber, Brad M, Xing Huang, Philippe Jorion, Terrance Odean, and Christopher Schwarz, 2022, A (Sub) penny For Your Thoughts: Tracking Retail Investor Activity in TAQ, *Available at SSRN 4202874* .

Bartlett, Robert P, Justin McCrary, and Maureen O'Hara, 2022, A Fractional Solution to a Stock Market Mystery, *Available at SSRN* .

Battalio, Robert, and Craig W Holden, 2001, A Simple Model of Payment for Order Flow, Internalization, and Total Trading Cost, *Journal of Financial Markets* 4, 33–71.

Battalio, Robert, Robert Jennings, and Jamie Selway, 2001, The Relationship Among Market-making Revenue, Payment for Order Flow, and Trading Costs for Market Orders, *Journal of Financial Services Research* 19, 39–56.

Boehmer, Ekkehart, Charles M Jones, Xiaoyan Zhang, and Xinran Zhang, 2021, Tracking Retail Investor Activity, *Journal of Finance* 76, 2249–2305.

Bryzgalova, Svetlana, Anna Pavlova, and Taisiya Sikorskaya, 2022, Retail Trading in Options and the Rise of the Big Three Wholesalers, *Available at SSRN* .

Ernst, Thomas, and Chester S Spatt, 2022, Payment for Order Flow and Asset Choice, *NBER Working Paper 29883* .

Gensler, Gary, 2021, "Prepared Remarks at the Global Exchange and FinTech Conference", Speech: Prepared Remarks at the Global Exchange and FinTech Conference [Accessed: 2022 08 29].

Gensler, Gary, 2022, "Competition and the Two SECs:", Remarks Before the SIFMA Annual Meeting, Washington, D.C. [Accessed: 2022 11 01].

Hendershott, Terrence, Saad Ali Khan, and Ryan Riordan, 2022, Option Auctions, *Working Paper* .

Hu, Edwin, and Dermot Murphy, 2022, Competition for Retail Order Flow and Market Quality, *Available at SSRN 4070056* .

Jain, Pankaj K, Jared A Linna, and Thomas H McInish, 2021, An examination of the NYSE's retail liquidity program, *The Quarterly Review of Economics and Finance* 80, 367–373.

Jain, Pankaj K, Suchi Mishra, Shawn O'Donoghue, and Le Zhao, 2020, Trading Volume Shares and Market Quality in a Zero Commission World, *Available at SSRN 3741470* .

Kyle, Albert S, 1985, Continuous auctions and insider trading, *Econometrica: Journal of the Econometric Society* 1315–1335.

Menkveld, Albert J, Bart Zhou Yueshen, and Haoxiang Zhu, 2017, Shades of darkness: A pecking order of trading venues, *Journal of Financial Economics* 124, 503–534.

Parlour, Christine A., and Uday Rajan, 2003, Payment for Order Flow, *Journal of Financial Economics* 68, 379–411.

Schwarz, Christopher, Brad M Barber, Xing Huang, Philippe Jorion, and Terrance Odean, 2022, The 'Actual Retail Price' of Equity Trades, *Available at SSRN 4189239* .

Schwenk-Nebbe, Sander, 2021, The Participant Timestamp: Get The Most Out Of TAQ Data, *Available at SSRN 3984827* .

SEC, 2013, Self-Regulatory Organizations; The NASDAQ Stock Market LLC; Order Granting Approval to Proposed Rule Change, as Modified by Amendment No. 1, to Establish the Retail Price Improvement Program on a Pilot Basis until 12 Months from the Date of Implementation.

Appendix A. Proofs

Proof of Proposition 1

Consider any $i \in \{1, 2...N\}$ and $(x, y) \in \left[-\frac{1}{2}, \frac{1}{2}\right] \times \left[-\frac{1}{2}, \frac{1}{2}\right]$, let

$$G(x|y) = \text{Prob}\left[\min_{-i} y_{-i} \leq x | y_i = y\right]$$

and

$$g(x|y) = \frac{dG(x|y)}{dx}.$$

It's easy to show

$$G(x|y) = 1 - \left(\frac{1}{2} - x\right)^{N-1},$$

$$g(x|y) = (N-1)\left(\frac{1}{2} - x\right)^{N-2}.$$

Let

$$\begin{aligned} v(x, y) &= \mathbb{E}\left[c_i | \min_{-i} y_{-i} = x, y_i = y\right] \\ &= c_0 + c_1 \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} y_j | \min_{-i} y_{-i} = x, y_i = y\right] + c_2 \mathbb{E}\left[y_i | \min_{-i} y_{-i} = x, y_i = y\right] \\ &= c_0 + \left(\frac{c_1}{N} + c_2\right) y + \frac{c_1}{N} x + c_1 \frac{N-2}{N} \frac{1}{2}\left(\frac{1}{2} + x\right) \\ &= \left(c_0 + c_1 \frac{N-2}{4N}\right) + \frac{c_1}{2} x + \left(\frac{c_1}{N} + c_2\right) y. \end{aligned}$$

We focus on symmetric equilibrium. Suppose market maker i's opponents use a continuous, increasing strategy $\beta(y)$ at time 0. And suppose market maker i observes signal y and reports signal z, its expected profit is

$$\begin{aligned} U_i(z, y) &= \text{Prob}\left(z \leq \min_{-i} y_{-i} | y\right)\left[\beta(z) - \mathbb{E}\left(c | z \leq \min_{-i} y_{-i}, y_i = y\right)\right] \\ &= [1 - G(z|y)]\left[\beta(z) - \frac{1}{1 - G(z|y)} \int_z^{\frac{1}{2}} g(x|y)\, v(x, y)\, dx\right] \\ &= [1 - G(z|y)]\,\beta(z) - \int_z^{\frac{1}{2}} g(x|y)\, v(x, y)\, dx. \end{aligned}$$

Market maker i's optimization condition (necessary condition) is

$$\left.\frac{\partial U_i(z,y)}{\partial z}\right|_{z=y}=0.$$

This is

$$-g(y|y)\beta(y)+(1-G(y|y))\beta'(y)+g(y|y)v(y|y)=0.$$

Simplifying the condition, we get

$$-\beta(y)+\left(\frac{1-G(y|y)}{g(y|y)}\right)\beta'(y)+v(y|y)=0. \tag{A1}$$

Let's conjecture that $\beta(y)$ is linear, i.e., there exist k_0, k_1 such that

$$\beta(y)=k_0+k_1y.$$

Substitute this into A1, we have

$$-(k_0+k_1y)+\frac{\frac{1}{2}-y}{N-1}k_1+\left(c_0+c_1\frac{N-2}{4N}\right)+\frac{c_1}{2}y+\left(\frac{c_1}{N}+c_2\right)y=0.$$

Then k_0,k_1 are solved by

$$-k_0+\frac{\frac{1}{2}k_1}{N-1}+c_0+c_1\frac{N-2}{4N}=0$$

$$-k_1-\frac{k_1}{N-1}+\frac{c_1}{2}+\frac{c_1}{N}+c_2=0$$

Then we get

$$k_1=\frac{N-1}{N}\left(\frac{c_1}{2}\frac{N+2}{N}+c_2\right),$$

$$k_0=c_0+\frac{c_1}{4N}\left(N-1+\frac{2}{N}\right)+\frac{c_2}{2N}.$$

It's easy to check that

$$\left.\frac{\partial U_i(z,y)}{\partial z}\right|_{z=y}=0$$

is also the sufficient condition in the optimization problem in this linear equilibrium.

Proof of Lemma 1

First let's introduce the random variable

$$r = \min_i y_i \in \left[-\frac{1}{2}, \frac{1}{2}\right]$$

and its CDF

$$H(r) = 1 - \left(\frac{1}{2} - r\right)^N$$

and PDF

$$h(r) = N\left(\frac{1}{2} - r\right)^{N-1}.$$

Then the total expected profit of market makers is

$$\begin{aligned}
W_M^{OBO} =& \mathbb{E}\left\{\mathbb{E}\left[k_0 + k_1 r - c_0 - c_1 \frac{1}{N}\sum_{j=1}^{N} y_j - c_2 r | \min_i y_i = r\right]\right\} \\
=& \mathbb{E}\left\{k_0 + k_1 r - c_0 - c_1\left(\frac{r + (N-1)\left(\frac{1}{2} + r\right)\frac{1}{2}}{N}\right) - c_2 r\right\} \\
=& \mathbb{E}\left\{\frac{\left(\frac{1}{2} - r\right)(c_1 + c_2 N)}{N^2}\right\} \\
=& \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\left(\frac{1}{2} - r\right)(c_1 + c_2 N)}{N^2} N\left(\frac{1}{2} - r\right)^{N-1} dr \\
=& \frac{1}{N+1}\left(\frac{c_1}{N} + c_2\right).
\end{aligned}$$

The expected total profit of investors W_I is

$$\begin{aligned}
W_I^{OBO} &= -\mathbb{E}\left[k_0 + k_1 r | \min_i y_i = r\right] \\
&= -\int_{-\frac{1}{2}}^{\frac{1}{2}} (k_0 + k_1 r) N\left(\frac{1}{2} - r\right)^{N-1} dr \\
&= -\left[c_0 + \frac{1}{N(N+1)} c_1 - \frac{N-3}{2(N+1)} c_2\right]
\end{aligned}$$

and the total welfare W_{total} is

$$\begin{aligned} W_{total}^{OBO} &= \mathbb{E}\left\{\mathbb{E}\left[-c_0 - c_1\frac{1}{N}\sum_{j=1}^{N} y_j - c_2 r|\min_i y_i = r\right]\right\} \\ &= W_M^{OBO} + W_I^{OBO} \\ &= -\left(c_0 - \frac{N-1}{N+1}\frac{c_2}{2}\right). \end{aligned}$$

Proof of Proposition 2

Both $W_{total}^{BR} < W_{total}^{OBO}$ and $W_M^{BR} < W_M^{OBO}$ are obvious. And

$$\begin{aligned} & W_I^{BR} < W_I^{OBO} \\ \Longleftrightarrow & \frac{1}{N(N+1)}c_1 - \frac{N-3}{2(N+1)}c_2 < 0 \\ \Longleftrightarrow & \frac{c_1}{c_2} < \frac{N(N-3)}{2}. \end{aligned}$$