Would Order-By-Order Auctions Be Competitive?*

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

June 6, 2023

Abstract

We model two methods of executing segregated retail orders: broker's routing, whereby brokers allocate orders using market maker's overall performance, and order-by-order auctions, where market makers bid on individual orders, a recent SEC proposal. Order-by-order auctions improve market maker allocative efficiency, but face a winner's curse reducing retail investor welfare, particularly when liquidity is limited. Additional market participants competing for retail orders fail to improve total efficiency and investor welfare when entrants possess information superior to incumbent wholesalers. Existing Retail Liquidity Programs empirically suggest order-by-order auctions would attract few bidders in less liquid stocks and low-liquidity periods.

*For helpful comments and feedback we are thankful to: Yashar Barardehi, Robert Battalio, Svetlana Bryzgalova, Eric Budish, Larry Glosten, Terrence Hendershott, Pete Kyle, Dmitry Livdan, Mark Loewenstein, Josh Mollner, Alberto Teguia, and Bart Zhou Yueshen, along with seminar and conference participants at Arizona State University, Carnegie Mellon–Penn State–Pittsburgh joint conference, Eastern Finance Association, HEC (Paris), Imperial College, London Business School, University of Maryland, the Microstructure Exchange, New York University Stern Market Microstructure Conference, Securities and Exchange Commission Economics Conference, Society for Financial Studies Cavalcade, and numerous anonymous industry participants.

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

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

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

I. Introduction

Retail order flow in U.S. equities is segregated, with retail brokers routing almost all their retail customer orders directly to market makers. These market makers assume a best execution obligation once they receive the order, whether they privately internalize trades off-exchange, or fill the retail orders from liquidity sourced from other sources, like exchanges or alternative trading systems (such as "dark pools"). Retail trades are attractive to market makers, either due to lower adverse selection, as in Battalio and Holden (2001), or due to their trades being less correlated, as in Baldauf, Mollner, and Yueshen (2022). In both cases, market makers provide retail investors with better prices than the exchanges. Recently, the SEC has proposed a change in market structure with the goal of potentially increasing competition for retail orders.[1] While previous academic work has explored *whether* retail flow should be segregated and its value, once segregated, the question of *how* segregated 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 order-by-order auctions. Our broker's routing model closely resembles the current market structure, with retail brokers determining where to route each order to maximize execution quality. While retail brokers use recent competing market maker performance to inform routing decisions, they do not communicate with any market maker prior to routing each individual order. Our order-by-order auction models the SEC's proposed Rule 615, which would mandate auctions for retail trades. These auctions would only be available to retail market orders, where any market participant could bid on each individual order.

We evaluate both models with a focus on inventory cost and competition. In our model, a broker receives an order from a retail investor and chooses one market maker to execute the order. Executing the order incurs (marginal) inventory costs for market makers. We assume that each market maker i receives a private liquidity signal y_i right before the retail broker routes the order and that inventory cost is affected by both the market maker's private signal and the average signal of all market makers. Intuitively, each private signal can be

[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).

thought of as the inventory position of market maker i, with a market maker's willingness to trade depending both on his private inventory and the aggregate liquidity of all market makers. To obtain the order, market participants submit their spreads simultaneously to the broker, and the one with the lowest bid obtains the order. The key difference between broker's routing and order-by-order auctions is the market participants' information set when they submit their spreads. Under an order-by-order auction, each market maker i submits the spread after observing the realization of y_i, while under broker's routing, the spread is submitted when each market maker i only observes a noisy signal about y_i. We solve the market equilibrium under both trading mechanisms and then identify differences in welfare distribution, inventory management, and order allocation efficiency that arise under each of the two market structures.

In the broker's routing setting, each market maker i only observes a noisy version of its liquidity signal y_i, when submitting the spread. This is motivated by our observation of broker's routing that when market makers compete for order flows, they are effectively competing for future order flows, and thus do not have accurate information about the cost of executing the orders. 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 about individual orders. Brokers route to a market maker, who must accept the order. In practice, evaluation of trades is done on a periodic (e.g., daily, weekly, or monthly) basis, where market makers strive to secure future order flows by enhancing the current execution quality. Nevertheless, when competing by improving current execution quality, they do not observe the actual inventory costs associated with orders that they will obtain in the future. We obtain a symmetric equilibrium for broker's routing, where the bid (spread) strategy of this equilibrium is monotone in the noisy signal. The broker routes the order to the market maker who submits the lowest spread. This delivers a highly competitive outcome with relatively low expected market maker profits, because market makers are less informationally heterogeneous and their bidding strategies are less dependent on their observed noisy signals. While the broker's routing setting delivers strong competition, the lack of communication on any specific trade means that a trade may be routed to a market maker who has high ex-post inventory cost, leading to inefficient order allocation and inventory management.

In the order-by-order auction model, in contrast, each market maker i bids after observing its true liquidity signal y_i. This is motivated by the SEC proposal that all retail orders are auctioned order-by-order, and if a market maker wins the auction, it has to execute the order immediately. When market makers compete for individual retail orders, they already have more accurate information about the marginal inventory cost of executing the order. In the auction, market makers' symmetric equilibrium bid (spread) is increasing in their private signals y_i, and thus in equilibrium, the participant with the lowest realized inventory cost will always win the auction with the most aggressive bid, leading to first-best allocative efficiency. The common-value nature of the auction, however, creates a winner's curse problem; whichever participant wins the auction learns that all other participants had higher cost signals. Consequently, market participants bid conservatively in the auction, and thus they earn more profit from the auction due to the strategic concern.

The welfare implications of switching from broker's routing to order-by-order auctions are thus determined by the trade-off between more efficient inventory management and the higher rent earned by market makers due to concerns about the winner's curse. As previously noted, our inventory cost has two components: the common-value component, which depends on the average liquidity signals of all market makers, and the independent private-value component, which depends only on the liquidity signal of the individual market maker. The private value component of cost underlies the considerations of allocative efficiency: order-by-order auctions ensure the lowest-cost market maker always obtains the order, which creates a welfare gain shared by investors and market makers. The public component, which is the same for all market makers, does not influence allocative efficiency, but it does lead to a more severe winner's curse problem under order-by-order auctions, because all market makers observe more accurate signals about true liquidity. As a result, all market makers will bid more conservatively, leading to larger rents for market makers and worse welfare for retail investors. In summary, while order-by-order auctions offer higher allocative efficiency than broker's routing, they produce *less* competition than broker's routing.

Based on the above intuition, both total welfare and market maker's welfare increase when switching from broker's routing to order-by-order auctions. The change in investor welfare depends on the trade-off between the gain from more efficient order allocation and more rent earned by market makers due to the winner's curse. In stocks where the common-value component is a larger share of the inventory cost, like small stocks or times with low liquidity, order-by-order auctions deliver worse welfare for retail investors than the current system of broker's routing. For these stocks, the potential allocative efficiency gains of the auction are smaller than the welfare loss to retail investors from the greater winner's curse, and associated less aggressive bidding, of market participants.

We then extend our baseline model to include institutional investors, as a key objective of the SEC proposal is to enable institutional traders to trade directly with retail investors in auctions. While institutional traders can increase the number of bidders in an auction, we consider the friction in which they also have superior information about the fundamental value of the asset. Incumbent wholesalers, who have no information about the fundamental value, respond by bidding more cautiously in the auction, which leads to less competition. As a result, the overall welfare of retail investors can further decline in the switch to order-by-order auctions. Moreover, we also find a market segmentation result due to asymmetric information. When information asymmetries between institutional investors and incumbent wholesalers are sufficiently severe, only institutional investors will effectively compete for high-quality (low-cost) orders, while all market participants compete for low-quality (high-cost) orders. This leads to heterogeneous impacts of switching to order-by-order auctions on orders with different qualities.

Finally, we model the impact of order-by-order auctions on the cross-sectional liquidity of stocks. Under our model of broker's routing, market makers compete (or commit to a spread) before observing the characteristics of each order. Brokers can evaluate wholesalers based on their performance across all orders, including stocks with different sizes and liquidity levels, allowing for cross-subsidization. In equilibrium, wholesalers may incur losses when trading small, high-cost stocks, which are compensated by profits from trading large, low-cost stocks. This results in smoothed trading costs across stocks. However, switching to order-by-order auctions can substantially decrease market makers' incentives to trade small stocks, leading to more variation in spreads among stocks.

The competitiveness of broker's routing is more important when there are few potential bidders, while allocative efficiency is more important when there are many bidders. While order-by-order auctions only exist as a proposal, we empirically evaluate an existing analogue, Retail Liquidity Programs (RLP), to gain insight into bidding behavior that might arise under the proposed auctions. Both RLPs and the proposed auctions allow segmentation of retail flow, and both implement first-price sealed-bid auctions. Exchange RLPs allow market participants to provide liquidity to retail orders at will by posting hidden limit orders which are only accessible by retail investor orders. When there is at least one round lot (100 shares) of RLP interest, exchanges disseminate an indicative flag in the market data indicating the presence of RLP liquidity, though not revealing the exact size or price of the order. If multiple participants post in an RLP, the participant with the most aggressively priced order will have first priority for any incoming retail market order, mirroring the potential competitiveness and allocative efficiency of an order-by-order auction. Unlike the broker's routing system, where market makers must accept any flow the broker routes to them, posting limit orders in a RLP is entirely voluntary: there may be many market participants posting limit orders, or none at all.

As several industry participants have highlighted the parallels between RLPs and the proposed auctions (e.g., Bishop (2022)), we empirically investigate the RLPs, and the extent to which they can inform potential bidding participation in the proposed auctions. Quote activity in RLPs is surprisingly high, with active quotes 25% of the trading day for stocks in the Russell 1000. Volumes, however, are small, with RLP trades comprising less than 0.5% of total equity trading volume. RLP usage is voluntary, while the use of order-by-order auctions would be mandatory, so we investigate underlying reasons for low RLP usage. First, we find that price improvement in RLPs is, on average, less than 10% of the spread, while Dyhrberg, Shkilko, and Werner (2022) estimate that wholesalers give an average spread-improvement of 26%. Second, we find that RLP quote interest tends to occur during times when order imbalances are low, consistent with the voluntary nature of RLP liquidity provision, with this at-will bidding shared with the proposed auctions. Third, we find that, under the pecking order theory of Menkveld, Yueshen, and Zhu (2017), RLPs would rank as high-cost high-immediacy venues in the pecking order of venues, with market makers already sourcing liquidity from them to the extent that liquidity is available. Price impacts of trades in RLP programs are more sensitive to volatility; when volatility is 1% higher, exchange sub-penny trades have a price impact ten basis points higher, while off-exchange trades have a price impact only two basis points higher.

While five exchanges operate RLPs, the RLP operated by IEX has a special feature: quotes in the IEX RLP can only be posted at the mid-quote. As a result, any order interest in the RLP reveals a specific price (mid-quote) available for retail trades. A key assumption in the SEC economic analysis of Rule 615 is the level of interest of institutional traders in trading with retail, with the SEC offering suggestive evidence from CAT data on the prevalence of hidden mid-quote liquidity. We use public SIP data to show that the IEX RLP has two-sided liquidity less than 2% of the trading day, suggesting potentially low interest from institutional traders. With a proprietary sample of retail trades from a large retail broker, we show that retail investors obtain mid-quote pricing far more often when the IEX RLP has an active matching trade (e.g., a retail buy order when the RLP has a bid to sell at mid-quote). We also show that when the IEX RLP only has same-direction liquidity (e.g., a retail buy order when the RLP only has a quote to buy at mid-quote), retail investors obtain worse prices than when there is no RLP interest at all, consistent with the adverse selection from opportunistic institutional investors that we model.

The SEC notion of an order-by-order auction seeks to "instill greater competition for retail market orders." Under the current system of broker's routing, each order is sent to a single market maker with no pre-trade communication, where competition is measured by aggregate execution quality. Switching to an order-by-order auction offers a tempting increase in allocative efficiency, as the market participant with the lowest inventory cost always wins an auction. But this comes with a drawback, as the participant who wins by outbidding all competitors with less optimistic signals suffers the auction winner's curse. Participants scale back their bids, and obtain increased welfare in the order-by-order system. Retail investor welfare increases if each individual auction attracts many bidders; for auctions which may attract few bidders, like orders in small stocks or during periods of market volatility, the switch to order-by-order auctions may decrease retail investor welfare.

II. Prior Literature and Contribution

Several papers consider whether retail segmentation is optimal as a market design. Easley, Kiefer, and O'Hara (1996) suggest PFOF is related to skimming trades with lower adverse selection, Battalio and Holden (2001) argue retail investors have lower adverse selection, and Baldauf et al. (2022) argue retail investors are less correlated. Segregating retail flow allows better prices for retail traders, but different mechanisms for segregating retail flow are not explored. Motivated by the recent SEC call for order-by-order competition, our paper provides a theoretical analysis into two possible methods of executing segmented 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 due to lower competition.

Several studies, including Eaton, Green, Roseman, and Wu (2022) and Parlour and Rajan (2003), examine how retail participants themselves impact market liquidity. Our analysis focuses solely on the execution quality obtained by marketable retail orders, as these are the orders which would be subject to the SEC order-by-order proposal. Hu and Murphy (2022), Jain, Mishra, O'Donoghue, and Zhao (2020), and Schwarz, Barber, Huang, Jorion, and Odean (2022) all explore variation in execution quality among brokers. Market makers can offer two possible forms of superior prices: PFOF (payments from market makers to brokers) and price improvement (payments from market makers directly to retail customers). Brokers may or may not pass on the total extent of PFOF revenue back to their customers in the form of lower commissions, as documented in Battalio, Jennings, and Selway (2001), while Schwarz et al. (2022) and Battalio and Jennings (2022) highlight that brokers prioritize execution quality, even along dimensions not reflected in SEC 605 reports. Our welfare analysis focuses on the extent to which wholesalers compete for order flow, rather than the extent to which brokers rebate any PFOF received back to retail traders.

Liquidity varies considerably in the cross-section of stocks. Corwin and Coughenour (2008) argue specialists allocate attention to more liquid stocks during times of market stress, while Foley, Liu, Malinova, Park, and Shkilko (2020) show how tying DMM assignments in large and small stocks can lead to substantial increases in liquidity for small stocks with little to no observed harm for large stocks. In an extension to our model, we show how broker's routing can enable a similar cross-subsidization, which is not possible under order-by-order auctions.

Previous studies (Bernhardt and Hughson (1997) and Biais, Martimort, and Rochet (2000)) show that market makers can earn positive profits when competing for orders. Bernhardt and Hughson (1997) emphasize the importance of order splitting in the duopoly case, while Biais et al. (2000) study common-value auctions where multiple market makers compete for an informed order. In both papers, the key friction is the asymmetric information from the liquidity demand side, which refers to informed traders. Our study also predicts that market makers will earn positive profits in both the broker's routing and order-by-order auction settings. However, in contrast to the previous studies, there is no asymmetric information from the liquidity demand side in our model since retail orders are typically uninformed. In our model, market makers receive private signals about their inventory position, which weakens competition and ensures positive profits in equilibrium. Additionally, we extend our study to institutional traders who can privately obtain signals about asset quality and compete for order flows, as suggested by the SEC. We show that the additional adverse selection on the liquidity supply side may exacerbate market inefficiency, leading to a novel market segmentation prediction.

At the core of our analysis theoretically is the comparison between ex-ante and ex-post competition. As we demonstrate, it is not inherent that ex-post competition is stronger than ex-ante competition. In our setting greater heterogeneity ex-post actually undercuts the degree of competition compared to ex-ante competition. The mechanism design literature also offers a related perspective by highlighting the value of ex-ante pre-commitment for designing mechanisms (e.g., Myerson (1979); Harris and Townsend (1981)).

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. Another empirical analogue is in bond markets, where Hendershott, Livdan, and Schürhoff (2021) highlight a winner's curse in bond auctions; consistent with our model extension to opportunistic traders, which shows that additional bidders do not always improve outcomes, due to the winner's curse.

Our empirical analysis focuses on Retail Liquidity Programs (RLPs) offered by several exchanges. Aramian and Comerton-Forde (2023) describe RLPs in Europe, while Jain, Linna, and McInish (2021) provide an overview of the NYSE Retail Liquidity Program in 2015. Five U.S. exchanges now operate RLPs, and we analyze current RLP data through the lens of learning about potential number of bidders expected under the SEC's proposed order-by-order auctions. RLPs provide a competitive process for both traditional market makers and institutional investors to enter limit orders which offer potential price improvement to retail trades, and therefore offer insight into the potential number of bidders that order-by-order auctions could expect. The auctions may at times attract no bidders at all, and Battalio and Jennings (2023) look at the potential for the NBBO to move during the auction under the SEC's current proposed auction time of 100 to 300 milliseconds.

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 identical market makers indexed by $i \in \{1, 2, ...N\}$. Our broker's objective is to minimize the bid-ask spread paid by the investor.[2]

At time 0, the broker receives one unit sell order from the investor, and sends it to a market maker to execute the order by the end of time 0.[3] We assume that the retail investor is trading only for liquidity reasons, so there is no information about asset value contained in the direction of the order. If market maker i executes the order, it has to hold the position until time 1 which incurs (marginal) inventory cost ζ_i. The structure of ζ_i is specified later in this section. Let s_i be the half bid-ask spread offered by market maker i, then the profit that market maker i receives at time 1 is

$$s_i - \zeta_i.$$

We consider a tractable framework with linear equilibrium in the literature of common-value auctions (Klemperer (1999), Menezes and Monteiro (2004)). At time 0, each market maker i receives an i.i.d private liquidity shock y_i. For simplicity, we assume that y_i is drawn from a uniform distribution $U[-\frac{1}{2}, \frac{1}{2}]$. The inventory cost ζ_i has the following structure

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

where c_0, c_1 and c_2 are positive constants. Since each market maker can only observe his own liquidity shock, the inventory cost ζ_i is not fully observed by market maker i. The cost function consists of three components. The first term c_0 is the unconditional expected inventory cost of executing the order, which is the same for all market makers. The second

[2]Some brokers take payment for order flow (PFOF), which accrues to the broker rather than the retail investor. FINRA best-execution rules prevent considering PFOF in routing decisions. In practice, brokers set one PFOF rate across all wholesalers, so that conditional on setting this PFOF rate, they route to minimize spread. While we do not capture any trade-offs in setting a high PFOF rate (which may impact spreads), our broker's incentives are consistent with a broader goal of extracting welfare from wholesalers, via combined PFOF and spread.

[3]The direction of the order does not change our results.

term

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

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

$$c_2 y_i$$

represents ζ_i's exposure to the individual liquidity shock y_i, and the coefficient c_2 measures the sensitivity. In our model, the coefficients (c_0, c_1, c_2) are exogenous, and are determined by stock characteristics. For example, a stock that is about to announce earnings may have a very high c_1, with market makers very concerned about aggregate inventory imbalances. In contrast, a tick-constrained stock with low informational asymmetries may have a very low c_1 value, with market makers not very concerned about aggregate inventories.

A. *Order-by-order Auction*

First, we consider a hypothetical order-by-order auction mechanism. We model the order-by-order auction as a common-value auction. In order-by-order auctions, each market maker i submits the spread s_i after privately observing the realization of signal y_i at time 0, and thus it can choose its spread strategy according to its assessment of inventory cost. The broker observes spreads offered by all market makers, and sends the order to the winner with the lowest spread at the end of time 0.[4] At time 1, all players collect their payoffs. We focus on symmetric equilibria such that all market makers choose the same strategy.

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

$$s = k_0 + k_1 y,$$

[4]If more than one market maker submits the lowest spread, then the winner is chosen randomly among those who submit the lowest spread.

and both k_0 and k_1 are solved in equilibrium. The following proposition summarizes our results.

Proposition 1. *In the model of order-by-order auctions, there exists a linear symmetric equilibrium in which the spread submitted by market maker* $i \in \{1, 2, ...N\}$ *is*

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

where

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

and

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

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

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

This slope k_1 is increasing in both c_1 and c_2 (see Plot 1). When c_1 and c_2 are increasing, market maker i's inventory cost is more sensitive to its private signal y_i. As a result, its spread s_i will also be more sensitive to the private signal y_i. The constant term k_0 is an increasing function of all three constants c_0, c_1 and c_2. Intuitively, k_0 is increasing in c_0, as a higher expected inventory cost forces market makers to bid wider spreads. Furthermore, k_0 is also increasing in both c_1 and c_2. Note that k_0 is the offered spread when any market maker observes the average signal $y = 0$. When both c_1 and c_2 increase, the variation of inventory cost will be larger among market makers. As a result, the marginal cost (or the probability) of losing the bid from marginally increasing the spread is lower, which motivates the market maker to choose a higher spread. Intuitively, when market makers differ more ex-post from one another, they are 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 the lowest signal realization, and thus the lowest inventory cost. Order-by-order auctions, therefore, implement the first-best outcome in terms of efficient allocation of the retail order, as the retail order is always matched to the market maker with the lowest inventory cost.

Plot 2 illustrates the equilibrium strategy across varying numbers of market makers, denoted as N. It is evident that the equilibrium bidding is a (weakly) decreasing function of N. This trend arises because as the number of bidders (market makers) increases, the significance of the winner's curse concern becomes more prominent. Consequently, market makers submit lower spreads for any given signal. An interesting observation is that the equilibrium bid is independent of the number of market makers N when observing the highest signal, $y = \frac{1}{2}$. This result is very intuitive. Since the equilibrium bid is an increasing function of the signal, the only case that a market maker who observes the signal $y = \frac{1}{2}$ can obtain the order is when all market makers observe the same, highest signal $y = \frac{1}{2}$. In this case, the market is competitive and thus the equilibrium bid must be

$$s = c_0 + c_1 \frac{1}{2} + c_2 \frac{1}{2}$$

which is independent of the number of market makers.

B. *Broker's Routing*

In this section, we consider the market equilibrium under broker's routing. In our model, we highlight the key difference between broker's routing and order-by-order auctions as market makers' different information sets when choosing spreads. Specifically, under broker's routing, market makers do not receive accurate signals about inventory cost when they compete. In practice, brokers and market makers establish long-term relationships. Market-maker performance is evaluated in the aggregate, but not order-by-order, and market makers cannot choose when to accept order flow from the broker; when a broker sends an order, they must fulfill it either by internalizing the order, or paying taker fees to fill the order at an exchange. Focusing on this key difference, we model broker's routing by assuming that each market maker i only receives a noisy signal about y_i when submitting the spread s_i, and that they are not able to adjust their spreads ex-post. Formally speaking, there is an additional stage, time -1, at which each market maker i receives a signal w_i. The signal w_i has the following structure. With probability p_0, $w_i = y_i$; and with probability $1 - p_0$, w_i is drawn from a uniform distribution $U\left[-\frac{1}{2}, \frac{1}{2}\right]$ which is independent of all other variables in the model. Each market maker i does not know whether w_i equals to y_i or not, and only understands that $w_i = y_i$ with probability p_0. Under broker's routing, all market makers submit their spreads at the end of time -1.

We still focus on symmetric equilibria in this case. In the model of broker's routing, each market maker i only observes the imperfect signal w_i when they submit their spread $t_i(w_i)$. Similar to our discussion on order-by-order auctions, we conjecture (and verify later) that there exists a linear symmetric equilibria where all market makers choose the same strategy $t(w)$, where

$$t = K_0 + K_1 w.$$

We refer readers to the appendix for more details and only present the equilibrium result.

Proposition 2. *In the model of broker's routing, there exists a linear symmetric equilibrium in which the spread submitted by market maker $i \in \{1, 2, ...N\}$ is*

$$t\left(w_i\right) = K_0 + K_1 w_i,$$

where

$$K_0 = c_0 + \frac{p_0}{4N^2}\left[\left(3 + N^2 - p_0 - Np_0\right)c_1 + 2Nc_2\right]$$

and

$$K_1 = \frac{N-1}{N}\left(c_2 p_0 + \frac{2c_1 p_0}{N} + \frac{c_1\left(N-2\right)p_0}{2N} + \frac{c_1\left(1-p_0\right)p_0}{2N}\right).$$

While we model broker's routing as a form of auction, it can also be natural to consider quantity competition in broker's routing (Kyle (1985), Baldauf et al. (2022)). Our goal here is to set up a comparable benchmark for order-by-order auctions, which features price competition. We likewise consider price competition for the broker's routing system.

We highlight the key difference between order-by-order auctions and broker's routing as the different information environments in which they compete. In our model of broker's routing, if $p_0 = 1$, it becomes the model of order-by-order auctions. At the other extreme when $p_0 = 0$, market makers are homogeneously uninformed when they submit their spreads, as they have not yet observed their private signals. As a result, Bertrand competition obtains, and all market makers will earn zero expected profit in equilibrium. Therefore, the unique symmetric equilibrium spread in this case must be

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

for all $i \in \{1, 2, ...N\}$. Plot 3 illustrates how the equilibrium strategy changes with p_0 in this case.

Although all market makers earn a non-negative expected profit, their ex-post profit can be positive or negative, depending on the realized inventory cost. In other words, market makers will lose money on some trades. In contrast, the realized profit in order-by-order auctions must be non-negative for all market makers for all trades. Second, under broker's routing, the order will be obtained by the market maker with the lowest signal w_i, who may not be the one with the lowest inventory cost as w_i is just a noisy signal of y_i. As a result, a welfare loss incurs due to inefficient inventory management in equilibrium. We present a more detailed welfare analysis in the next subsection.

C. Welfare Analysis: Order-by-order Auction vs. Broker's Routing

In our model, the retail order is always executed, but inventory cost and equilibrium spreads differ between order-by-order auctions and broker's routing. Let $j \in \{OBO, BR\}$, we denote W_M^j, W_I^j and W_{total}^j as the market makers' expected profit, the investor's expected profit and the total welfare under order-by-order auctions and broker's routing, respectively:

1. The expected total profit of all market makers W_M^j: the expected equilibrium spread minus the incurred inventory cost;

2. The expected total profit of the retail investor W_I^j: the expected negative equilibrium spread;

3. The total welfare W_{total}^j: the expected negative incurred inventory cost, which is $W_{total}^j = W_M^j + W_I^j$.

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

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

The investor's expected profit is

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

and the total welfare is

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

Total welfare is the sum of the welfare of market makers and the retail investor: $W_{total}^{OBO} = W_M^{OBO} + W_I^{OBO}$. Based on our equilibrium results, we obtain the following Lemma.

Lemma 1. *Under order-by-order auctions, the welfare outcomes of the equilibrium characterized by Proposition 1 are*

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

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

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

Order-by-order auctions implement the first-best allocation, and the total welfare is

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

It is clear that W_{total}^{OBO} is decreasing in the expected inventory cost c_0. Furthermore, W_{total}^{OBO} is increasing in c_2, because c_2 determines the variation of inventory cost among all market makers. When c_2 is higher, the expected lowest inventory cost will be lower, and thus the total welfare is higher. W^{OBO}_{total} is independent of c_1, because the aggregate component

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

in the inventory cost always has zero mean. That is, the aggregate contribution of the common-value component in the inventory cost is always zero, no matter how large is c_1.

Costs incurred from c_0 are borne exclusively by the investor and do not factor into market makers' welfare W^{OBO}_M. An increase in c_1 leads to an improvement in market makers' welfare as each market maker's private information becomes more relevant in the calculation of inventory costs, resulting in a more diverse bidding strategy. This, in turn, leads to market makers earning a higher information rent from the auction. Since c_1 has no impact on total welfare, when c_1 increases, investor welfare will decrease due to market makers earning higher information rents. On the other hand, although an increase in c_2 always leads to an increase in market makers' welfare, the retail investor's welfare increases only when the number of market makers $N > 3$. First, a higher c_2 implies a better inventory management outcome and total welfare as discussed earlier, which is beneficial and is a positive effect for the investor. However, a higher c_2 also implies that market makers are ex-post more different and their bidding strategies will be more diverse, leading to positive rent earned by market makers and is a negative effect for investors. It turns out that the positive effect (strictly) dominates only when $N > 3$, because the rent earned by market makers becomes less crucial when there are more market makers. As a result, the investor can benefit from a higher c_2 only when $N > 3$.

Under broker's routing, all welfare calculations are similar, except that market makers only observe noisy signals about y_i. For simplicity of exposition, we skip the intermediate steps and only present the final results.

Lemma 2. *Under broker's routing, the welfare outcomes of the equilibrium characterized by Proposition 2 are*

$$W^{BR}_M = \frac{p_0 \left(2c_1 - p_0 c_1 + N c_2\right)}{N \left(1 + N\right)},$$

$$W^{BR}_I = -\left[c_0 + p_0 \frac{2\left(2 - p_0\right) c_1 - \left(N - 3\right) N c_2}{2N \left(1 + N\right)}\right],$$

$$W_{total}^{BR} = W_M^{BR} + W_I^{BR} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

Having solved for welfare outcomes in the model of broker's routing and order-by-order auctions, we next compare welfare between the two mechanisms in the following proposition.

Proposition 3. *$W_{total}^{BR} < W_{total}^{OBO}$; $W_M^{BR} < W_M^{OBO}$; $W_I^{BR} < W_I^{OBO}$ if and only if $N(N-3)\frac{c_2}{c_1} > 2(1-p_0)$.*

Proposition 3 is a direct result of Lemma 1 and Lemma 2. Note that the only aggregate welfare loss in our model is from inefficient inventory management. The improvement of total welfare

$$W_{total}^{OBO} - W_{total}^{BR} = (1-p_0)\frac{N-1}{N+1}\frac{c_2}{2}$$

is increasing in N and decreasing in p_0. Intuitively, when the ex-ante signal is less noisy (p_0 is higher), the order is more likely to be obtained by the market maker with the lowest inventory cost, and thus the welfare loss from inefficient order allocation will be lower. The magnitude of welfare improvement also depends on the number of market makers. When there are more maker makers, the first-best allocation will be more efficient as their inventory costs are not perfectly correlated. Order-by-order auctions implement the first-best outcome, while the outcome of broker's routing depends on the precision of the ex-ante signal, and is less sensitive to the number of market makers. Consequently, the welfare improvement from broker's routing to order-by-order auctions is higher when there are more market makers. Finally, it is worth noting that the welfare improvement is also increasing in c_2. This is because, when c_2 is higher, the market makers' ex-post inventory cost will be more diverse. This means that the first-best allocation, which minimizes the inventory cost, will be more efficient. However, while order-by-order auctions implement the first-best allocation, broker routing does not, and the total welfare under broker's routing is less sensitive to c_2 compared to that under order-by-order auctions. As a result, the total welfare difference between these two mechanisms is larger when c_2 is higher.

Market makers' welfare is higher under order-by-order auctions, and the difference is:

$$W_M^{OBO} - W_M^{BR} = \frac{1}{N+1}\left((1-p_0)^2\frac{c_1}{N} + (1-p_0)c_2\right).$$

The difference in market maker welfare $W_M^{OBO} - W_M^{BR}$ is increasing in c_1 and c_2. When c_1 and c_2 are higher, the private signals that market makers observe become more important in their inventory cost. Market makers are, effectively, more different ex-post. The ex-post heterogeneity generates the positive expected profit they earn under order-by-order auctions. The difference $W_M^{OBO} - W_M^{BR}$ is also increasing in $(1-p_0)$, as the precision of the noisy signal in the model of broker's routing determines the competitiveness of the market. When the signal is noisier, i.e., $(1-p_0)$ is higher, the market under broker's routing is more competitive, resulting in a lower expected profit for market makers, and thus the difference of market makers' welfare under these two mechanisms becomes larger.

The contrast between W_I^{BR} and W_I^{OBO} depends on the levels of c_1, c_2 and the number of market makers N, reflecting the trade-off between more efficient inventory management and higher rent earned by market makers driven by their heterogeneity in information and concerns about winner's curse. The winner's curse concern arises from the common-value component in the inventory cost, represented by c_1 in the inventory cost. When market makers bid, they must consider that winning is conditional on other market makers observing relatively high-cost signals, which can lead to market makers bidding more conservatively in equilibrium. This effect tends to increase the (expected) equilibrium spread, which ultimately hurts the investor and is more significant under order-by-order auctions. On the other hand, the impact of the private-value component in the inventory cost, represented by c_2, depends on the number of market makers. When switching from broker's routing to order-by-order auctions, the investor's welfare will be more sensitive to c_2, but the sign of the effect depends on the number of market makers, as illustrated earlier. When $N > 3$, a higher c_2 leads to higher welfare for the investors, and the investor can benefit from switching to order-by-order auctions when c_2 is high enough. A special case is when $c_1 \to 0$, the comparison between W_I^{BR} and W_I^{OBO} will only depend on the number of market makers N, but not the level of c_2. The following Lemma highlights this case.

Lemma 3. *When $c_1 \to 0$, the inventory costs of market makers become mutually independent, and $W_I^{BR} < W_I^{OBO}$ if and only if $N > 3$.*

Proof. This is a direct result of Proposition 3. □

A direct implication of Proposition 3 is that switching to order-by-order auctions has heterogeneous impacts on stocks with different inventory cost structures. Compared to small and illiquid stocks, large and liquid stocks usually can be executed by the market makers and thus rely less on the interdealer market.[5] As a result, $\frac{c_2}{c_1}$ will be relatively larger for large liquid stocks and the smaller stock is more likely to breach the threshold $\frac{2(1-p_0)}{N(N-3)}$ when $N > 3$. Order-by-order auctions are therefore more likely to harm the investor's welfare in small illiquid stocks compared to large liquid stocks.

We do not directly model the endogenous entry of market makers, but our model gives implications for how market competition and liquidity provision change welfare outcomes in partial equilibrium analysis. Proposition 3 implies that when the number of market makers N is small, the investor's welfare is likely lower upon switching to order-by-order auctions. Here N measures the number of active market makers who provide liquidity. During time periods when market makers are not willing to provide liquidity (for example, due to market uncertainty or high inventory cost), our model predicts that switching to order-by-order auctions will be more likely to hurt investors. Since investor protection may be more important during market distress, our result highlights the unintended negative effect of order-by-order auctions during time periods when liquidity provision is limited.

D. The Role of Institutional Traders

In the proposal released by the SEC, the entry of institutional traders has been highlighted as a key feature of order-by-order auctions.[6] The SEC hopes that, relative to the current broker's routing system, order-by-order auctions will allow institutional traders to increase competition for retail trades.[7] This hope, however, ignores the fact that institutional traders usually have superior information about asset quality compared to wholesalers (e.g., Glosten and Milgrom (1985)). Allowing institutional traders to compete for retail orders may increase information asymmetry among bidders in order-by-order auctions, and lead to a less efficient equilibrium outcome. In this section, we develop a model to analyze this extension and demonstrate that the entry of institutional traders can introduce more competition, which benefits retail investors, however, it also may introduce more adverse selection, which can

[5] See a microfoundation of this intuition in the Internet Appendix.

[6] As the SEC chairman Gary Gensler mentioned, "...individual investors don't necessarily get the best prices that they could get if institutional investors, like pension funds, could systematically and directly compete for their orders." Gensler (2021)

[7] A "wholesaler is often chosen by a formula that depends on past execution quality of the wholesaler, its relationship with the broker-dealer, and other factors. In addition, the bilateral nature of the wholesaler business model not only restricts contemporaneous competition among wholesalers, it also restricts opportunities for other market participants" Securities and Exchange Commission (2022).

negatively impact market outcomes. The overall impact of the entry of institutional traders depends on the level of information asymmetry in the market.

To extend our model to include institutional traders, we make two (minimal) changes in the baseline model. First, apart from the N wholesalers[8] who always provide market-making service, there are $N_0 \geq 2$ institutional traders who also can provide liquidity but only in order-by-order auctions. This is consistent with the market design suggested in the SEC proposal, in which institutional traders are absent in the current broker's routing system, but can be active and provide more competition in order-by-order auctions. We assume that institutional traders $i \in \{1, 2, ...N_0\}$ also receive i.i.d private signals y_i at time 0, which follows a uniform distribution $U\left[-\frac{1}{2}, \frac{1}{2}\right]$. The private signal y_i plays a similar role as that for wholesalers, as discussed later in the model.

Second, we consider the following (new) inventory cost structure

$$\tilde{\zeta}_i = \tilde{c}_0 + c_1 \frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{N}} y_j + c_2 y_i, \tag{1}$$

where $\tilde{c}_0$ is a random variable that can be $c_0 - \delta_c$ or $c_0 + \delta_c$[9] with equal probabilities, and $\tilde{N}$ is the number of active market makers for this order.[10] In broker's routing, we only have wholesalers competing for retail orders, so our equilibrium features $\tilde{N} = N$; and in order-by-order auctions, both wholesalers and institutional traders can compete for retail orders, so $\tilde{N}$ depends on the number of active market makers in this case. Note that $\mathbb{E}(\tilde{c}_0) = c_0$, that is, the unconditional expectation of inventory cost remains the same in this extension.

Institutional traders have an information advantage over wholesalers. Specifically, all institutional traders can observe the realization of $\tilde{c}_0$ at time 0, while wholesalers only know the distribution of $\tilde{c}_0$. This implies that when competing for retail orders, institutional traders can condition their bids on the realization of $\tilde{c}_0$, while wholesalers can only use distributional information of $\tilde{c}_0$.[11] This information asymmetry captures the notion that institutional traders are more informed about the characteristics of assets traded, market

[8]These are the market makers in our baseline model.

[9]Without loss of generality, we assume $\delta_c \geq 0$.

[10]"Active market makers" means those who have a non-zero probability of obtaining the order in equilibrium.

[11]We can also interpret $\pm\delta_c$ as private information of asset quality, and keep the inventory cost structure unchanged. This will not change our model outcomes.

conditions, or future price movement.

We first consider the market equilibrium in the broker's routing system. Since institutional investors are absent in broker's routing, the only difference between this extension and our baseline model is the structure of inventory cost. The additional randomness in the inventory cost (1) has no impact on market equilibrium, because all wholesalers are risk neutral and thus only care about the expectation of the $\tilde{c}_0$. Recall that $\mathbb{E}(\tilde{c}_0) = c_0$, which is the same as in the baseline model.

Proposition 4. *With inventory cost structure* (1)*, under broker's routing, the equilibrium bidding strategies and welfare outcomes are the same as characterized by Proposition 2 and Lemma 2.*

The market equilibrium is unchanged under broker's routing, thus we view it as a suitable benchmark of the (new) model with institutional traders. However, the equilibrium does change under order-by-order auctions due to the entry of institutional traders. First, consider the case when $\delta_c = 0$, and thus institutional traders have no informational advantage compared to wholesalers. In this case, the only effect is enhanced competition in order-by-order auctions, which is a direct result of the increased number of bidders competing for the retail order. With Proposition 1 obtained in our baseline model, to obtain the new market equilibrium, we simply replace the number of bidders N in the baseline model with $(N + N_0)$, because institutional traders are ex-ante identical to wholesalers in this special case. The following Proposition characterizes this equilibrium.

Proposition 5. *When there are N_0 institutional traders and $\delta_c = 0$, under order-by-order auctions, there exists a linear symmetric equilibrium in which the spread submitted by wholesaler or institutional trader i is*

$$\tilde{s}_i(y_i) = \tilde{k}_0 + \tilde{k}_1 y_i$$

where

$$\tilde{k}_0 = c_0 + \frac{c_1}{4(N+N_0)}\left(N + N_0 - 1 + \frac{2}{N+N_0}\right) + \frac{c_2}{2(N+N_0)}$$

and

$$\tilde{k}_1 = \frac{N+N_0-1}{N+N_0}\left(\frac{c_1}{2}\frac{N+N_0+2}{N+N_0} + c_2\right).$$

We consider the total welfare $\tilde{W}^{OBO}_{total}$, the investor's welfare $\tilde{W}^{OBO}_{I}$, and wholesalers' welfare $\tilde{W}^{OBO}_{W}$. Institutional traders' welfare $\tilde{W}^{OBO}_{IT}$ satisfies

$$\tilde{W}^{OBO}_{IT} = \tilde{W}^{OBO}_{total} - \tilde{W}^{OBO}_{I} - \tilde{W}^{OBO}_{W}.$$

Denote the total welfare, the investor's welfare, and wholesalers' welfare under broker's routing as $\tilde{W}^{BR}_{total}$, $\tilde{W}^{BR}_{I}$, and $\tilde{W}^{BR}_{W}$, respectively. We then compare welfare outcomes under broker's routing and order-by-order auctions in this extension.

Proposition 6. *When there are N_0 institutional traders and $\delta_c = 0$, we have the following results on welfare comparison:*

1. $\tilde{W}^{BR}_{total} < \tilde{W}^{OBO}_{total}$;
2. $\tilde{W}^{BR}_{W} < \tilde{W}^{OBO}_{W}$ *if and only if* $\frac{N(N+1)}{(N+N_0)(N+N_0+1)} > p_0$ *and* $\frac{c_2}{c_1} > -\frac{1}{N+N_0}\frac{\frac{N(N+1)}{(N+N_0)(N+N_0+1)} - p_0\frac{(N+N_0)(2-p_0)}{N}}{\frac{N(N+1)}{(N+N_0)(N+N_0+1)} - p_0}$;
3. $\tilde{W}^{BR}_{I} < \tilde{W}^{OBO}_{I}$ *if and only if* $\frac{c_2}{c_1} > \frac{\frac{1}{(N+N_0)(1+N+N_0)} - \frac{p_0(2-p_0)}{N(N+1)}}{\frac{N+N_0-3}{2(N+N_0+1)} - \frac{p_0(N-3)}{2(N+1)}}$.

When institutional traders provide liquidity in order-by-order auctions but have no informational advantage, the total welfare unambiguously improves when switching from broker's routing to order-by-order auctions. Since the order is always obtained by the market maker with the lowest ex-post inventory cost under order-by-order auctions and their inventory costs are not perfectly correlated, having institutional traders in order-by-order auctions always makes the order allocation more efficient. The effect on investor's welfare is ambiguous, which is higher under order-by-order auctions if and only if $\frac{c_2}{c_1}$ is greater than a threshold

$$\frac{\frac{1}{(N+N_0)(1+N+N_0)} - \frac{p_0(2-p_0)}{N(N+1)}}{\frac{N+N_0-3}{2(N+N_0+1)} - \frac{p_0(N-3)}{2(N+1)}},$$

qualitatively similar to the Proposition 3 in the baseline model. If

$$\frac{1}{(N+N_0)(1+N+N_0)} - \frac{p_0(2-p_0)}{N(N+1)} < 0, \tag{2}$$

the threshold is always negative, and the investor's welfare always improves under order-by-order auctions, irrespective of the level of $\frac{c_2}{c_1}$. Condition (2) concerns the number of new institutional traders providing liquidity under order-by-order auctions. Under the joint assumption that institutional traders have no informational advantage when they compete for the retail order (i.e., when $\delta_c = 0$) and that order-by-order auctions can attract sufficiently many institutional traders, investors will unambiguously benefit from switching to order-by-order auctions, as the benefit of efficient inventory management will dominate any decreases in competition. This is precisely the intuition motivating the SEC's proposal on order-by-order auctions, and our above results highlight the underlying assumptions required for it to hold. After switching to order-by-order auctions, the wholesalers' welfare is increasing if and only if two conditions are satisfied. First, the number of new institutional investors N_0 has to be low enough, i.e.,

$$\frac{N\left(N+1\right)}{\left(N+N_0\right)\left(N+N_0+1\right)} > p_0.$$

Unconditionally, all wholesalers and institutional traders can obtain the order with equal probabilities. Additional institutional investors decreases the probability a wholesaler obtains the order and mechanically decreases wholesalers' welfare; with sufficiently many institutional investors, wholesalers are worse off relative to broker's routing.

Second, $\frac{c_2}{c_1}$ must exceed the threshold:

$$\frac{c_2}{c_1} > \frac{-\frac{1}{N+N_0}\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)}{\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}}. \tag{3}$$

Note that wholesalers' welfare unambiguously improves from broker's routing to order-by-order auctions in our baseline model. However, with the entry of institutional traders, the wholesalers' welfare increases only when

$$\frac{c_2}{c_1}$$

is sufficiently high, because the entry of new institutional traders also decreases the total welfare of all market makers (wholesalers and institutional traders). When $\frac{c_2}{c_1}$ is sufficiently high, the market makers' inventory cost will be more heterogeneous ex-post, creating more information rent for them. Therefore wholesalers' welfare is higher under order-by-order auctions only when $\frac{c_2}{c_1}$ is sufficiently high.

We next consider the case when institutional investors have at least some informational advantage, that is when $\delta_c > 0$. We still focus on symmetric linear equilibria where all institutional traders choose the same linear strategy and all wholesalers choose the same linear strategy. When $\delta_c > 0$, institutional traders can condition their bids on the realization of $\tilde{c}_0$. Intuitively, when observing $\tilde{c}_0 = c_0 - \delta_c$, institutional traders will submit lower bids, and when $\tilde{c}_0 = c_0 + \delta_c$, they will submit higher bids. In contrast, wholesalers cannot condition their spreads on the realizations of $\tilde{c}_0$, but just the distributional information c_0. Wholesalers are more likely to win auctions when $\tilde{c}_0 = c_0 + \delta_c$ than when $\tilde{c}_0 = c_0 - \delta_c$, as institutional traders will bid more aggressively in the latter case. This leads to adverse selection for wholesalers, as they are more likely to win auctions when $\tilde{c}_0 > E(\tilde{c}_0)$. A winner's curse argument implies that wholesalers will submit more conservative bids in equilibrium. When δ_c is sufficiently large, the winner's curse concern becomes so severe, such that all wholesalers will be completely out of competition for high-quality (low-cost) stocks, and can only obtain the retail order when $\tilde{c}_0 = c_0 + \delta_c$. Consequently, when $\tilde{c}_0 = c_0 - \delta_c$, institutional traders will face no competition from wholesalers, which can reduce retail investor welfare. The following proposition formalizes this intuition:

Proposition 7. *Let* $\tilde{s}^-(y; \delta_c)$ *and* $\tilde{s}^+(y; \delta_c)$ *be two bidding strategies, where*

$$\tilde{s}^-(y; \delta_c) = \tilde{k}_0^-(\delta_c) + \tilde{k}_1^-(\delta_c)\, y_i$$

with

$$\tilde{k}_0^-(\delta_c) = c_0 - \delta_c + \frac{c_1}{4N_0}\left(N_0 - 1 + \frac{2}{N_0}\right) + \frac{c_2}{2N_0}$$
$$\tilde{k}_1^-(\delta_c) = \frac{N_0 - 1}{N_0}\left(\frac{c_1}{2}\frac{N_0 + 2}{N_0} + c_2\right),$$

and

$$\tilde{s}^+(y; \delta_c) = \tilde{k}_0^+(\delta_c) + \tilde{k}_1^+(\delta_c)\, y$$

with

$$\tilde{k}_0^+(\delta_c) = c_0 + \delta_c + \frac{c_1}{4(N + N_0)}\left(N + N_0 - 1 + \frac{2}{N + N_0}\right) + \frac{c_2}{2(N + N_0)}$$
$$\tilde{k}_1^+(\delta_c) = \frac{N + N_0 - 1}{N + N_0}\left(\frac{c_1}{2}\frac{N + N_0 + 2}{N + N_0} + c_2\right).$$

When there are N_0 *institutional traders, there exists a threshold* $\underline{\delta} > 0$*, such that when* $\delta_c > \underline{\delta}$*,* *there exists an equilibrium of order-by-order auctions in which*

1. *the wholesalers always choose bidding strategy* $\tilde{s}^+(y;\delta_c)$*;*
2. *institutional traders choose bidding strategy* $\tilde{s}^+(y;\delta_c)$ *when observing* $c_0+\delta_c$ *and* $\tilde{s}^-(y;\delta_c)$ *when observing* $c_0-\delta_c$*.*

The threshold $\underline{\delta}$ satisfies the following condition

$$\tilde{k}_0^-(\underline{\delta})+\tilde{k}_1^-(\underline{\delta})\frac{1}{2}<\tilde{k}_0^+(\underline{\delta})-\tilde{k}_1^+(\underline{\delta})\frac{1}{2}.$$

This implies that when the true state is $\tilde{c}_0=c_0-\delta_c$, the highest possible spread offered by institutional traders is still lower than the lowest possible spread offered by wholesalers, and thus wholesalers will never obtain the order in this case, irrespective of their signal realizations. When the true state is $\tilde{c}_0=c_0+\delta_c$, wholesalers and institutional traders will choose the symmetric bidding strategy $\tilde{s}^+(y;\delta_c)$, and thus all players will obtain the order with equal probabilities in this case.

If we interpret the random variable $\tilde{c}_0$ as the heterogeneous quality of stocks, then in equilibrium, only institutional traders effectively compete for retail orders of high-quality stocks, while all market makers compete for orders of low-quality stocks. The market for low-quality stocks becomes more competitive due to an increase in the number of bidders, while the market for high-quality stocks may become less competitive as institutional traders are the only effective bidders. The presence of adverse selection can weaken competition and potentially harm total welfare, as illustrated by our following proposition.

Proposition 8. *When there are* N_0 *institutional traders and* $\delta_c>\underline{\delta}$*, we have the following results on welfare comparison:*

1. $\tilde{W}_{total}^{BR}<\tilde{W}_{total}^{OBO}$ *if and only if* $N_0>\underline{N}_0$*, where* $\underline{N}_0$ *is a constant solved in appendix by* (I.A.3.4)*;*
2. $\tilde{W}_W^{BR}<\tilde{W}_W^{OBO}$ *if and only if* $p_0<\frac{1}{2}\frac{N}{N+N_0}\frac{1+N}{N+N_0+1}$ *and* $\frac{c_2}{c_1}>\frac{-\frac{1}{N+N_0}\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)}{\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}}$*;*
3. $\tilde{W}_I^{BR}<\tilde{W}_I^{OBO}$ *if and only if* $\left(1-p_0+\frac{4p_0}{N+1}-\frac{2}{N+N_0+1}-\frac{2}{N_0+1}\right)\frac{c_2}{c_1}>\frac{1}{(N+N_0)(N+N_0+1)}+\frac{1}{N_0(N_0+1)}-\frac{2p_0(2-p_0)}{N(1+N)}$.

The comparison of total welfare between broker's routing and order-by-order auctions is complicated by the presence of adverse selection. The transition from broker's routing to order-by-order auctions results in an improvement in total welfare only when there is a sufficient number of institutional traders providing liquidity. However, the absence of wholesalers can result in a decrease in market competitiveness for high-quality stocks, leading to an inefficient outcome for their trading. While low-quality stocks may experience an increase in market competitiveness, this gain may not be enough to offset the welfare loss from the trading of high-quality stocks. Dyhrberg et al. (2022), using SEC 605 reports, show that while total institutional trading in S&P 500 stocks exceeds retail investor volume, institutional trading volume drops off for less liquid stocks. For the least liquid tercile of U.S. equities, they find that retail investor trading volume is four times (median) to seven times (average) as large as institutional trading volume, raising questions about whether order-by-order auctions would garner enough institutional interest to clear the thresholds set in Proposition 8.

In accordance with Proposition 6, the welfare effects on both the investor and wholesalers are similar. Both parties are likely to benefit from stocks with a high $\frac{c_2}{c_1}$. As previously noted in our baseline model (and appendix), stocks with a high $\frac{c_2}{c_1}$ tend to be large and highly liquid, leading to welfare losses for small, illiquid stocks.[12]

E. Heterogeneous Stocks and Cross-subsidization

In our baseline model, we consider one unit order from a single stock, and the equilibrium and welfare outcomes depend on (c_0, c_1, c_2). In this section, we extend our baseline model to heterogeneous stocks with different characteristics (c_0, c_1, c_2). For order-by-order auctions, this extension is straightforward, as the stock characteristics (c_0, c_1, c_2) are publicly observable when market makers compete. Then the market equilibrium (spread, allocation, and welfare outcomes) can still be captured by our baseline model. The extension, however, is less straightforward for broker's routing. This is because broker's routing features the long-term relationship between brokers and market makers, and thus the competition among market makers happens before the order actually arrives and the order characteristics are observed. As a result, market outcomes among heterogeneous stocks under broker's routing

[12] This is also in line with the concerns expressed by practitioners, who generally believe that the transition to order-by-order auctions may negatively impact small and illiquid stocks, as noted in SIFMA (2022).

will be less differentiated compared to that under order-by-order auctions. Our model predicts that compared to order-by-order auctions, there is less variation in equilibrium spreads among stocks under broker's routing. Based on this observation, we can also highlight a cross-subsidization effect: under broker's routing, the equilibrium spreads of high-cost stocks are relatively low (compared to that under order-by-order auctions), while the equilibrium spreads of low-cost stocks are relatively high. This cross-subsidization effect implies that switching from broker's routing to order-by-order auctions not only changes the retail investors' welfare but also changes the welfare distribution when retail investors have different portfolio holdings.

To capture this idea, we consider a pool of orders characterized by a joint cumulative distribution $G(c_0, c_1, c_2)$, and the realization of (c_0, c_1, c_2) is independent of all other variables in the model. For simplicity, we assume that G has full support on $(0, \infty) \times (0, \infty) \times (0, \infty)$, and is continuously differentiable everywhere. We consider a model with the following timeline:

1. At time -1, the cumulative distribution function $G(c_0, c_1, c_2)$ becomes public information, and each market maker i observes his private noisy signal w_i;

2. At time 0, an order with characteristics (c_0, c_1, c_2) is drawn from distribution G, and each market maker i observes his private signal y_i. The broker then sends the order (c_0, c_1, c_2) to one market maker;

3. At time 1, all random variables are realized and all market participants collect their payoffs.

As we discussed in the baseline model, under broker's routing, market makers compete and submit their spreads at time -1, while under order-by-order competition, they submit their spreads at time 0.[13] We first introduce the following variables

$$\bar{c}_0 = \iiint c_0 dG\left(c_0, c_1, c_2\right) = \mathbb{E}(c_0),$$

$$\bar{c}_1 = \iiint c_1 dG\left(c_0, c_1, c_2\right) = \mathbb{E}(c_1),$$

$$\bar{c}_2 = \iiint c_2 dG\left(c_0, c_1, c_2\right) = \mathbb{E}(c_2).$$

Under order-by-order auctions, since order characteristics (c_0, c_1, c_2) are public, the equilibrium and welfare outcomes are the same as characterized by Proposition 1 and Lemma 1 in our baseline model.

Under broker's routing, since only distributional information G is available when market makers compete at time -1, the equilibrium strategy will only depend on the distributional information G but not the specific order characteristics (c_0, c_1, c_2). The new equilibrium of broker's routing is characterized by the following Proposition.

Proposition 9. *In the extension of heterogeneous stocks, under broker's routing, there exists an equilibrium in which every market maker who observes signal w chooses to submit spread*

$$\bar{T}\left(w\right) = \bar{K}_0 + \bar{K}_1 w$$

where

$$\bar{K}_0 = \bar{c}_0 + \frac{p_0}{4N^2}\left[\left(3 + N^2 - p_0 - Np_0\right)\bar{c}_1 + 2N\bar{c}_2\right]$$

[13]The literature on combinatorial auctions (e.g., Pekeč and Rothkopf (2003); De Vries and Vohra (2003)) considers auctions of multiple goods with non-additive values. In our model, the value derived from executing multiple retail orders is additive for market makers. Other papers (e.g., Adams and Yellen (1976); McAfee, McMillan, and Whinston (1989); Jehiel, Meyer-Ter-Vehn, and Moldovanu (2007)) have examined models of multi-object auctions with additive values, showing that bundling can sometimes outperform selling multiple goods separately. Given this context, a natural question in this extension is: can the investor benefit from an optimal design of order bundling? Suppose the broker can select a subset $B \subseteq A$, where A represents the set of all possible values of (c_0, c_1, c_2), and conduct two auctions. One auction includes orders with $(c_0, c_1, c_2) \in A$, while the other includes orders from $(c_0, c_1, c_2) \in A - B$. As we will see later, the investor's profit is linear in (the expected value of) (c_0, c_1, c_2), implying that neither separating nor bundling orders increases the investor's profit in our extension, and thus considering optimal bundling of orders is unnecessary in this extension of broker's routing.

and

$$\bar{K}_1 = \frac{N-1}{N}\left(\bar{c}_2 p_0 + \frac{2\bar{c}_1 p_0}{N} + \frac{\bar{c}_1 (N-2) p_0}{2N} + \frac{\bar{c}_1 (1-p_0) p_0}{2N}\right).$$

Since all market makers are risk neutral and the equilibrium is linear in the baseline model, we still obtain a linear equilibrium in this extension. Consider K_0 and K_1 in the baseline model as functions of (c_0, c_1, c_2), the equilibrium strategy in this extension satisfies

$$\bar{T}(w) = \mathbb{E}(t(w)) = \mathbb{E}(K_0 + K_1 w) = \bar{K}_0 + \bar{K}_1 w.$$

Then market makers choose an "average" bidding strategy in this extension. Note that both K_0 and K_1 are increasing functions of c_0, c_1 and c_2, this result implies that, compared to our baseline model results, the equilibrium spread in this extension is relatively low for stocks with high inventory cost characteristics, and high for stocks with low inventory cost characteristics.

The welfare impacts also are heterogeneous. To be specific, we consider the welfare outcomes for any specific order with characteristics (c_0, c_1, c_2). The following Lemma summarizes our results.

Lemma 4. *In the equilibrium characterized by Proposition 9, the investor's welfare* $\bar{W}^{BR}_{heter,I}$*, the total welfare* $\bar{W}^{BR}_{heter,total}$ *and market makers' welfare* $\bar{W}^{BR}_{heter,M}$ *are*

$$\bar{W}^{BR}_{heter,I} = -\left[\bar{c}_0 + p_0 \frac{2(2-p_0)\bar{c}_1 - (N-3) N \bar{c}_2}{2N(1+N)}\right],$$

$$\bar{W}^{BR}_{heter,total} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right),$$

$$\begin{aligned}\bar{W}^{BR}_{heter,M} &= \bar{W}^{BR}_{heter,total} - \bar{W}^{BR}_{heter,I} \\ &= (\bar{c}_0 - c_0) + \frac{p_0}{2(N+1)}\left[(N-1)c_2 - (N-3)\bar{c}_2\right] + p_0 \frac{(2-p_0)\bar{c}_1}{N(1+N)}.\end{aligned}$$

The total welfare in Lemma 4 is the same as that in the baseline model. Note that in our model, the total welfare is only determined by inventory cost but not the equilibrium spread, as the spread is just a transfer between market makers and the investor. Since the order is always obtained by the market maker with the lowest signal w_i, and introducing heterogeneity in stocks does not change allocative efficiency, we conclude that the total welfare is the same as that in the baseline model for any order (c_0, c_1, c_2) in this extension. However, the equilibrium spread does change. Specifically, now the investor's welfare (which is the negative expected equilibrium spread) becomes

$$-\left[\bar{c}_0 + p_0 \frac{2\left(2-p_0\right)\bar{c}_1 - \left(N-3\right)N\bar{c}_2}{2N\left(1+N\right)}\right]$$

which only depends on the average levels $(\bar{c}_0, \bar{c}_1, \bar{c}_2)$, but not order characteristics (c_0, c_1, c_2). Note that under order-by-order auctions, the investor's welfare is

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

which depends on order characteristics (c_0, c_1, c_2). Then investors will be worse off after switching to order-by-order auctions if for their orders, c_0 is high, c_1 is high, or c_2 is low (when $N > 3$). This highlights our cross-subsidization effect under broker's routing that market makers charge relatively low spreads for high-cost stocks and relatively high spreads for low-cost stocks. This cross-subsidization effect also implies that switching from broker's routing to order-by-order auctions may have unintended effects on retail investors' welfare distribution. For example, investors who mainly trade small, illiquid stocks with high average inventory cost c_0 will be worse off after switching to order-by-order auctions, while those who trade large, liquid stocks with low average inventory cost c_0 will be better off.

The market makers' welfare is

$$\left(\bar{c}_0 - c_0\right) + \frac{p_0}{2\left(N+1\right)}\left[\left(N-1\right)c_2 - \left(N-3\right)\bar{c}_2\right] + p_0\frac{\left(2-p_0\right)\bar{c}_1}{N\left(1+N\right)}$$

which depends on the difference between order characteristics (c_0, c_1, c_2) and the average levels $(\bar{c}_0, \bar{c}_1, \bar{c}_2)$. Under order-by-order auctions, the market makers' welfare

$$\frac{1}{N+1}\left(\frac{c_1}{N} + c_2\right)$$

is always positive. However, under broker's routing, marker makers make more profit from stocks with relatively low inventory cost, and incur loss from stocks with high inventory cost, and the welfare (or net profit) from executing a specific order may be negative. This is consistent with our observation that under the current broker's routing system, market makers sometimes lose by providing liquidity for small, illiquid stocks but they can make a profit from executing large liquid stocks. On average, they can make positive expected profit from market making. Our result implies that, after switching to order-by-order auctions, market makers will only submit spreads that are high enough such that they can earn a positive expected profit on every individual order.

IV. Suggestive Bidding in Retail Liquidity Programs

The trade-off between the greater allocative efficiency of order-by-order auctions and greater competition of broker's routing depends crucially on the number of bidders. The number of potential bidders who would participate in order-by-order auctions is unknown, but there are close existing analogues, retail liquidity programs (RLPs). Allison Bishop, co-founder of Proof Trading, draws the comparison to RLPs, noting, "exchanges already have ways for retail orders to be identified and treated specially by market makers, called retail liquidity programs (RLPs)... [which] can deliver a similar benefit to retail investors through order-by-order competition among market makers and institutional investors." – Bishop (2022). In approving exchange applications to offer RLPs, the SEC itself has frequently highlighted the same features that it has highlighted for proposing order-by-order competition, namely to increase the number of market participants interacting with retail orders. Given the comparisons drawn by industry participants between the auctions and existing RLPs, we empirically investigate the RLPs to gain insight into the potential bidding participation in the proposed order-by-order auctions.

Five exchanges have developed retail liquidity programs (RLPs). Any market participant can, through their broker, place a hidden limit order in the RLP, where this hidden limit order is only accessible to retail market orders. As a result, RLPs function as continuous first-price double auctions for any incoming retail market orders. Like the proposed order-by-order auctions, they are also sealed-bid auctions, as all RLP limit orders are hidden. Exchanges are allowed to disseminate a flag of interest, however, indicating whether there are at least 100 shares of RLP liquidity available on either the ask side, bid side, or both sides of the market.

We conduct two lines of inquiry into RLPs. First, we examine narrowly the IEX RLP to gain insight into the existing potential institutional interest in bidding against retail traders. While most RLPs allow sub-penny pricing in tenths of pennies, the IEX RLP only allows mid-quote pricing. We use this feature to evaluate independently the number of bidders predicted by the SEC economic analysis. Second, we examine broadly all five RLP programs, with the goal of learning about the cross-sectional characteristics of RLPs as well as the performance of an auction-like format for smaller or less liquid stocks.

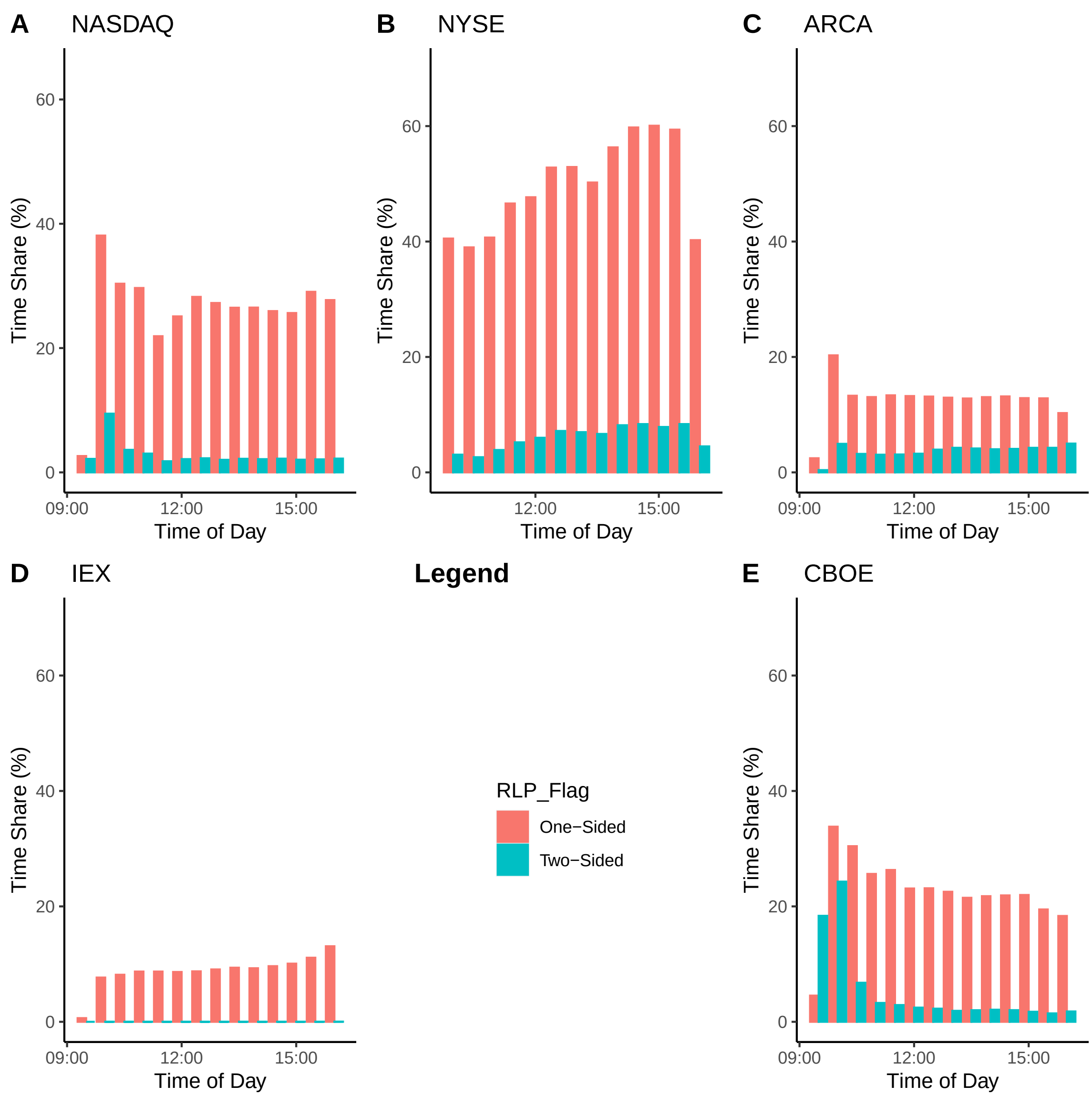
A. Data and Summary Statistics

We obtain NYSE TAQ (Trade and Quote) data from January 1, 2019 to May 30, 2022. We examine all securities in the Russell 3000 index, as well as the 100 most frequently traded ETFs, provided these securities are priced above $1 per share. Retail Liquidity Program (RLP) Indicators are distributed through the SIP, and are available in TAQ Data. As indicators may be disseminated even when an exchange's visible posted best bid or offer (BBO) is not at the official NBBO, we obtain retail indicator flags from the TAQ Quotes file.

A relatively high percentage of the trading day has at least some RLP interest. Figure 1 highlights variation across programs throughout the trading day. Across each time interval, the IEX RLP is active for a notably smaller percentage of time relative to any competing RLPs, as the IEX RLP requires orders to be placed at mid-quote, while competitor programs only require a minimum of 10 mils of improvement relative to the NBBO. The RLP flags do not display the exact amount of price improvement available, nor do they display any indication of the size available, with the flag only indicating whether there is at least one round lot.

While the percentage of the trading day with RLP interest is considerable, the volumes executed through RLP programs are diminutive. Exchange RLP trades comprise 0.3% of all ETF trading volume, and 0.1% of trading in Russell 1000 or Russell 2000 stocks. Detailed summary statistics on the volume and usage of RLPs are provided in Internet Appendix I.A.1.

Figure 1. Intra-day Time Share. We plot the average percent of time that the RLP Flag is active throughout the trading day on January 3, 2022. For each exchange, we divide the trading day into 30-minute intervals and calculate the average across stocks of the percentage of time for which the RLP Flag is active. One-sided liquidity is the percentage of time for which there is a quote on either the bid, the ask, or both, and therefore includes the time for which there is two-sided liquidity (i.e., a flag indicating RLP interest on both the bid and ask at the same time).



B. SEC Proposal And Current IEX RLP Usage

Proposition 8 of our model highlights that if order-by-order auctions fail to attract sufficient volume, retail welfare will be worse under order-by-order auctions. As a result, the estimates of institutional interest in trading with retail investors in a sealed-bid auction setting are highly consequential in evaluating the SEC proposal. The SEC's economic analysis of the Proposed Rule 615 suggests that under the new auction format, institutional traders would potentially bid up price improvement to mid-quote as often as 75% of the time.[14] While the SEC's analysis uses non-public CAT data, the IEX RLP offers an alternative method for estimation of the interest of institutions in trading with retail at mid-quote. The IEX RLP allows market participants to post limit orders priced at the mid-quote which are only available to retail investors.

Figure 1 shows that the IEX RLP has, on average, RLP interest less than 15% of the trading day. Furthermore, the IEX RLP has two-sided interest less than 2% of the trading day. Based on the SEC CAT estimates of non-displayed hidden mid-quote liquidity, there must be a large fraction of institutional traders who desire trading at midpoint, but do not currently post in the IEX RLP program. This raises a question: why do these institutions not use the IEX RLP, and what does that suggest for their interest in using order-by-order auctions?

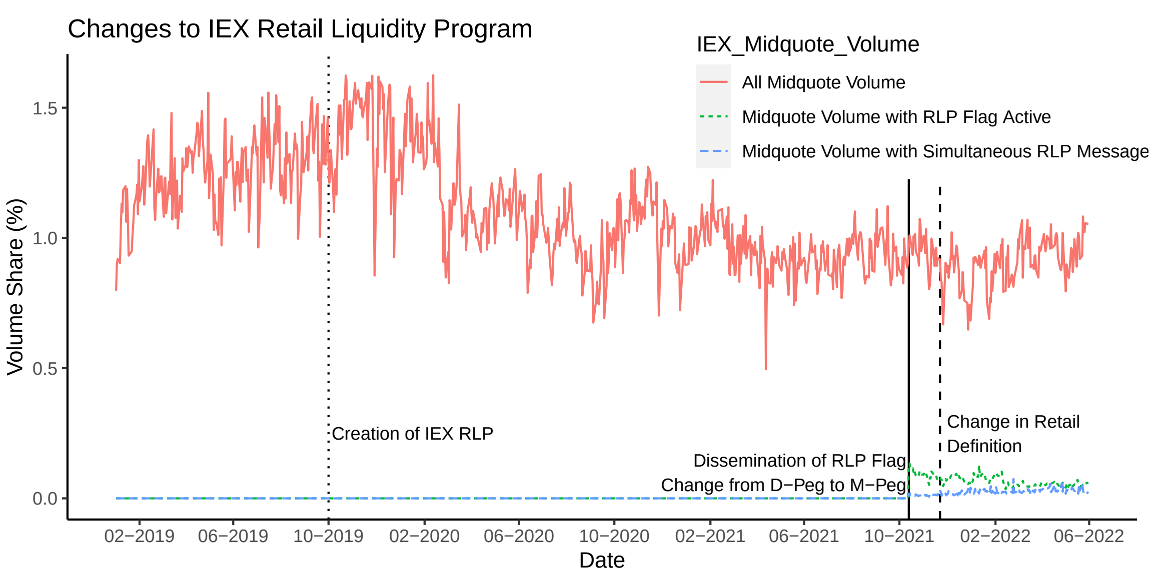
One natural possibility is that the institutional traders posting hidden mid-quote liquidity are not seeking to trade with retail investors, particularly as individual retail orders are often small. Another possibility is that institutional traders have concerns over information leakage. Prior rule changes in the IEX RLP allow us to further investigate this possibility, as the program historically allowed RLP orders but did not disseminate an indicative flag of RLP interest.

Figure 2 plots mid-quote trading volume at IEX; total hidden mid-quote orders at IEX (both RLP-only and traditional hidden orders) comprise around 1% to 1.5% of total U.S. equity trading volume, with no obvious change in this volume around the time the IEX RLP

[14] The SEC reports that "On average, 51% of the shares of individual investor marketable orders internalized by wholesalers are executed at prices less favorable than the NBBO midpoint ('Wholesaler Pct Exec Shares Worse Than Midpoint'). Out of these individual investors shares that were executed at prices less favorable than the midpoint, on average, 75% of these shares could have hypothetically executed at a better price against the non-displayed liquidity resting at the NBBO midpoint on exchanges and NMS Stock ATSs." Securities and Exchange Commission (2022)

is created on October 1, 2019. The IEX RLP began distributing an indicator message of RLP interest on October 13, 2021. Total mid-quote volume does not appreciably change in the month following the introduction of the IEX RLP indicator, which suggests that concerns over information leakage from the RLP indicator alone are not the primary determinant in posting RLP limit orders.

Figure 2. IEX Mid-quote Volume and Key RLP Rule Changes. The IEX Retail Liquidity Program was introduced on October 1, 2019, with only hidden discretionary midpoint-peg orders. On October 13, 2021, the Retail Liquidity Program changed the RLP order type to a midpoint-peg order and began dissemination of an indicator of whether there was RLP interest. On November 22, 2021, the requirement that retail investors submit no more than 390 orders per day was lifted. We plot total mid-quote volume on IEX (as a percentage of total equities trading volume) with the solid (red) line. We plot mid-quote volume which occurs during the time that the IEX RLP is active with the dotted (green) line. We plot the total mid-quote volume which occurs simultaneously with an RLP message with the dashed (blue) line.



Order-by-order auctions would also differ in the nature of their post-trade transparency. Current mid-quote liquidity offers no indication of trade direction: a mid-quote trade between an institutional investor and retail investor would not indicate which party was the buyer. In the proposed order-by-order auctions, the retail investor trade direction, and by extension any auction counter-party's trade direction, would be revealed. Analogously, RLPs can reveal direction if the RLP is one-sided, or if the retail market order consumes all liquidity in the RLP, resulting in the dissemination of a new RLP indicator.[15] As Figure 2 highlights, total mid-quote volume is over 10 times higher than RLP mid-quote volume, which may, in part, reflect concerns from institutional investors over post-trade transparency.[16]

Another possibility for the low usage of the IEX RLP is that posting in the IEX RLP does not facilitate trading with retail at mid-quote. To investigate this possibility, we obtain a sample of trades from a large retail brokerage. Our sample covers all retail orders for 100 shares or less on three days (March 13, 14, and 16, 2023) in a set of liquid symbols.[17] Figure 3 plots the cumulative distribution function of price improvement (measured as a percentage of the quoted bid-ask spread) of orders. Price improvement of 50% would correspond to execution at the NBBO mid-quote. When the IEX RLP is active, retail investors receive more price improvement: 75% of orders receive mid-quote or better when the flag is active, compared to 60% of orders when the flag is not active.

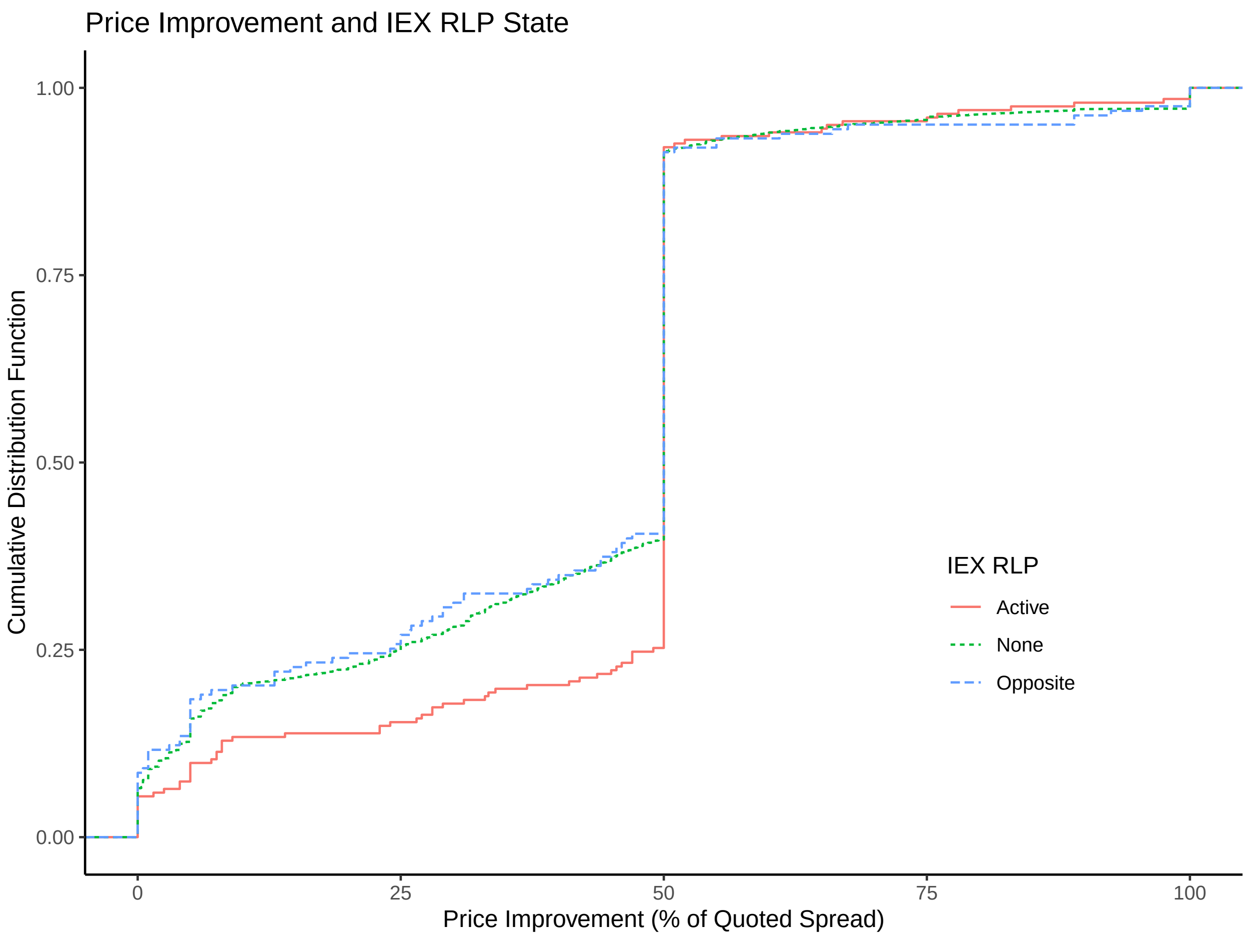
The share of retail orders receiving mid-quote when the IEX RLP is active is less than 100%. One possible explanation is that we do not observe the view of the wholesaler during the time the routing decision is made. In addition to the latency of the IEX speed bump, there is some latency between the wholesaler and a host of potential venues, and trying multiple venues may take milliseconds rather than microseconds. A second potential explanation is that not all wholesalers may connect to, or route to, the IEX RLP. While the IEX RLP flag indicates at least 100 shares are available at mid-quote, it is only available for a very short period of the trading day. Furthermore, we note that while retail orders are sometimes filled at a price worse than mid-quote when the IEX RLP is active, retail orders are also sometimes filled at prices *better* than mid-quote, even when the IEX RLP is not active.

[15]For example, if there is two-sided RLP liquidity, but there is a message of a mid-quote trade and a simultaneous message that the RLP changes from two-sided to bid-side interest only, market participants could infer that there was a retail order to sell. A standard (i.e., not RLP) hidden mid-quote order would avoid this information leakage.

[16]From the TAQ data, it is impossible to determine the exact portion of orders that are retail orders in the IEX RLP program, but we can estimate an upper and lower bound. For the upper bound, we count all mid-quote orders which occur when the IEX retail flag is active, though some of this volume may include non-retail mid-quote orders interacting with hidden mid-quote liquidity. For the lower bound, we measure mid-quote volume which has a simultaneous message update for the RLP program; this measures only retail orders which consume the available RLP liquidity (necessitating an updated RLP message), but will miss retail orders which do not consume all available RLP liquidity and therefore send no update message.

[17]The symbols are: AAPL, AMZN, CRDO, DIA, GOOG, IVV, IWM, JPM, MU, PPG, QQQ, SGH, SPY, V, VICR, and VOO.

Figure 3. Retail Trades and the IEX RLP. We obtain a sample trades of retail investors from a large retail brokerage. Our sample covers all retail orders for 100 shares or less on three days (March 13, 14, and 16, 2023) in a set of liquid symbols. We plot the empirical cumulative distribution function of price improvement (as a percentage of the quoted spread), received by retail orders. Active refers to either two-sided IEX RLP interest, bid interest at the time of a retail sell order, or ask interest at the time of a retail buy order. None refers to no active RLP flag. Same-direction refers to only RLP bid interest at the time of a retail sell order, or only RLP ask interest at the time of a retail buy order.



When the IEX RLP interest is one-sided and the same direction as retail investor trade, we find that retail investors receive less price improvement.[18] When the IEX RLP interest is

[18]By one-sided same direction interest, we refer to one-sided ask side interest when there is a retail sell order, or one-sided bid side interest when there is a retail buy order. Note that 'bid' and 'ask' refer to the limit order interest in buying or selling, and the price in the IEX RLP is always the NBBO mid-quote.

same-direction, retail investors obtain 5% less price improvement than when the IEX RLP has no interest at all. Proposition 6 in our model highlights that institutional traders can decrease retail investor welfare, as the presence of institutional traders can increase adverse selection, and lead wholesalers to bid more cautiously. While routing to the IEX RLP is voluntary, wholesalers currently appear to provide less price improvement when the IEX RLP has same-direction only interest. For example, retail buy orders obtain worse price improvement when there is IEX RLP interest on the bid side only compared to when there is no IEX RLP interest at all. This is consistent with the intuition of Proposition 7, whereby wholesalers adopt a more conservative bidding strategy in the presence of institutional investors, who possess information about securities.

C. RLP Program Usage and Market Conditions

Our model highlights how welfare gains or losses on order-by-order auctions compared to broker's routing hinges on both the number of bidders and the stock-specific characteristics. While predictions about the order-by-order auctions are inherently speculative, we examine the full set of five RLPs for indicative evidence as to the behavior of the cross-section of stocks in an auction trading mechanism.

Each of the exchange retail liquidity programs has the same at-will feature of liquidity provision as the order-by-order auctions, with liquidity-providing participants in the program under no obligation to guarantee execution of retail trades.[19] Consequently, exchange retail programs offer insight into the potential workings of an order-by-order model, where market participants are similarly under no obligation to participate for all orders. While many market makers may wish to provide liquidity for orders in large stocks during periods of low volatility, our model suggests this does not hold true for smaller or less liquid stocks. Motivated by this reasoning, we estimate the following regression.

[19]We note that the NYSE RLP does have a requirement that retail liquidity providers offer price-improving RLP 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 RLP orders are active.

REGRESSION 1: *For each asset i:*

$$RPI\_Volume\_Share_i = \alpha_0 + \alpha_1 Percent\_Time\_At\_Minimum\_Spread_i + \alpha_2 Market\_Cap_i + \alpha_3 Average\_Volume_i + \epsilon_{ijkt}$$

Results of Regression 1 are presented in Table I. We estimate volume as a percentage of total volume, and as a percentage of total sub-penny volume. Exchange RLP volume is considerably larger when assets spend a larger percentage of the day at the minimum bid-ask spread, is considerably larger for larger market-cap stocks, and is considerably larger for stocks with higher average trading volume. The low usage of RLPs in small, less-liquid stocks suggests a low number of natural potential bidders, consistent with Dyhrberg et al. (2022) estimates of total retail volume relative to institutional volume, and suggests retail welfare would decline under order-by-order auctions (Proposition 3).

Retail investors placing market orders may arrive at any time in the day, including during periods of stress. Even under the generous assumption that their orders are uncorrelated with aggregate institutional order flow, half their order flow would be in the same direction as aggregate institutional order flow.[20] To investigate the relationship between retail liquidity program volume and price movements, we estimate Regression 2 with volume and price impacts, with fixed effects for each stock and date, and present the results in Table II. 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.[21]

REGRESSION 2: *For each asset i on date t:*

$$VolumeShare_{it} = \alpha_0 + \alpha_1 Percent\_Time\_At\_Minimum\_Spread_{it} + \alpha_2 Volatility_{it} + \alpha_3 Average\_Volume_i + \alpha_4 Absolute\_Intraday\_Return_{it} + X + \epsilon_{it}$$

Retail liquidity programs offer less price improvement on average than off-exchange

[20] A second issue is whether there is any institutional interest at all. Dyhrberg et al. (2022) and Battalio, Jennings, Saglam, and Wu (2022) highlight that retail interest vastly exceeds institutional interest in less liquid securities.

[21] They are not a perfect analogue. Barardehi, Bernhardt, Da, and Warachka (2022) document that sub-penny trading may be driven not by the activity of retail investors, but the extent to which better improvement opportunities (such as mid-quote trading) are available.

Table I: Cross-Sectional Variation in Volume Shares. This table estimates Regression 1 with sub-penny volume, measured as a percentage of all volume, and as a percentage of sub-penny priced volume. For stock i on date t, $Percent\ Time\ At\ Minimum\ Spread_{it}$ measures the percentage of the trading day with a quoted bid-ask spread of one penny, $Volatility_{it}$ measures the standard deviation of 15-minute returns, $Market\ Cap$ measures the market capitalization of the stock in billions, and Average Volume measures the average trading volume in billions. Observations are at the stock (or ETF) level for the sample of securities described in Section IV.B.

	Dependent variable:	
	Percentage of All Volume	Percentage of Only Sub-penny Volume
	(1)	(2)
Market Cap	0.120***	0.926**
	(0.035)	(0.417)
Percent Time at Minimum Spread	0.001***	0.013***
	(0.0001)	(0.001)
Average Volume	0.041***	0.380***
	(0.004)	(0.042)
Constant	0.058***	0.931***
	(0.002)	(0.030)
Observations	2,590	2,590
R^2	0.159	0.108
Note:		*p<0.1; **p<0.05; ***p<0.01

wholesalers offer. Across our sample, the average sub-penny improvement received in RLP programs averages 11 hundredths of a cent. Across sub-penny off-exchange trades, the average sub-penny price improvement is 19 hundredths of a cent, though some of these trades may not be retail. Using SEC Rule 605 reports, Dyhrberg et al. (2022) estimate for retail trades in S&P 500 stocks, wholesalers offer, on average, 27% of the quoted half-spread as price improvement.[22] Under the pecking-order theory of Menkveld et al. (2017), investors

[22]Within RLP trades, we estimate sub-penny improvement at 11.0 hundredths of a cent across all trades, and 10.8 hundredths of a cent among S&P 500 trades. Given that around 80% of RLP volume occurs in

Table II: Panel Variation in Volume and Price Impact. This table estimates Regression 2 with sub-penny volume, expressed as a percentage of total trading volume, and price impact, measured in basis points 30 seconds after the trade. Observations are at the stock-day level. Volatility measures the standard deviation of 15-minute price changes. Percent time at Minimum Spread measures the percentage of time the stock spread is a single tick, while absolute intraday return measures the absolute value of the intraday return. We include a fixed effect for each stock and date, and cluster standard errors by stock and by date. Note that Price Impact cannot be calculated when there is zero volume, thus Columns 4, 5, and 6 differ in the number of stock-days with zero volume in each category.

Dependent Variable:	*Volume*			*Price Impact*		
Venue:	Exchange	Off	Off	Exchange	Off	Off
RLP Active:	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE
	(1)	(2)	(3)	(4)	(5)	(6)
Percent Time At Minimum Spread	0.001	-0.028^{***}	0.027^{***}	0.021	0.050	-0.072^{*}
	(0.010)	(0.005)	(0.009)	(0.021)	(0.038)	(0.037)
Volatility	6.592^{***}	-2.464^{***}	-4.128^{***}	9.520^{**}	2.281^{***}	1.221
	(0.510)	(0.261)	(0.471)	(4.582)	(0.381)	(0.745)
Absolute Intraday Return	1.324^{***}	-0.819^{***}	-0.505^{***}	-0.650	0.251	-0.162
	(0.041)	(0.034)	(0.032)	(0.662)	(0.306)	(0.127)
Observations	1,965,888	1,965,888	1,965,888	682,727	1,771,969	1,885,905
R^2	0.417	0.248	0.380	0.013	0.002	0.003
Residual Std. Error	38.068	27.755	31.363	392.416	403.481	438.930
Note:						*p<0.1; **p<0.05; ***p<0.01

target low-cost-low-immediacy venues first, and if they fail to find liquidity, they access higher-cost-higher-immediacy venues, particularly at times of market stress or volatility. Consistent with this prediction, we find that on-exchange trading in RLP programs is very sensitive to intra-day volatility, with larger volatility being associated with more exchange sub-penny trading. For off-exchange trading, the opposite is true, with larger volatility associated with less off-exchange sub-penny trading.[23]

S&P 500 stocks, the 27% of the quoted half-spread is the more appropriate benchmark.

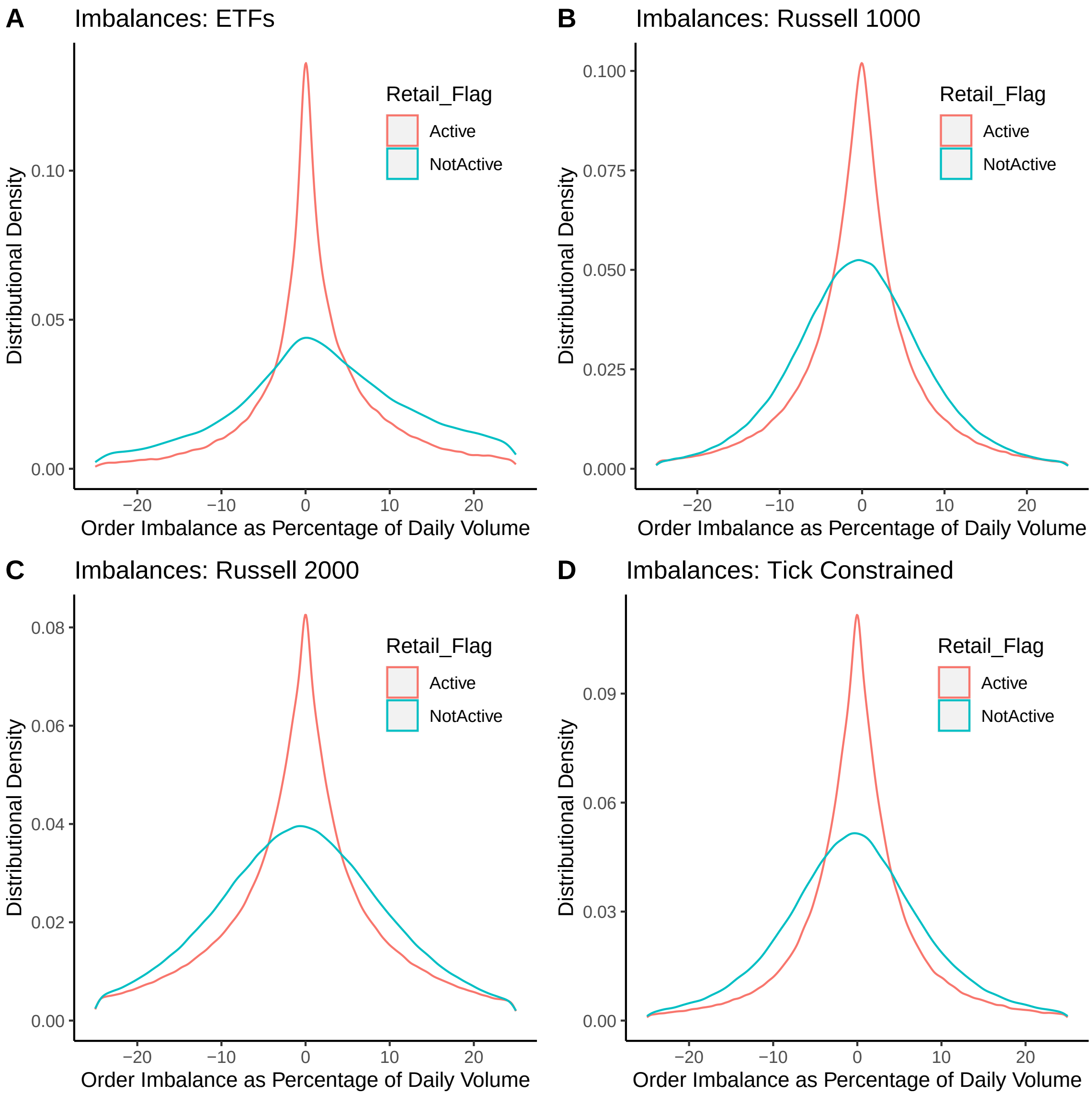
[23]Average price improvement given by RLP programs is less than the average improvement given by wholesalers or the average price improvement received by retail investors. Schwarz et al. (2022) estimate retail investors receive an average of 33% of the bid-ask spread. Routing to an RLP is voluntary, unlike the order-by-order auctions. While RLP programs currently reflect a high-cost trading venue in the pecking

In the economic analysis for the proposed Order-by-Order Competition Rule, the SEC argues that orders with lower price impact are equivalent to lower adverse selection risk: "Marketable orders internalized by wholesalers feature lower price impacts, i.e., have lower adverse selection risk." – Securities and Exchange Commission (2022). As one measure of adverse selection, we explore the pattern of order imbalances for on-exchange RLP trades and off-exchange sub-penny trades, depicted in Figure 4. 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,[24] but it is important to note that institutional investors may be eager to buy from retail investors at times, or sell to retail investors at times, but unlikely to want to stand ready to buy or sell to retail investors at any time on demand. Proposition 3 of our model highlights that when auctions have few participants, retail welfare is lower under auctions than the current system of broker's routing. Empirically, the level of participation in existing RLPs is negatively correlated with imbalances, suggesting analogous order-by-order auctions would have few bidders during periods of illiquidity.

order (where "cost" reflects the low level of price improvement given), this position may change if routing to the auction becomes mandatory.

[24]See SEC (2013).

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 ±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 at least 50% of the trading day for at least one-third of the trading days in our sample.



V. Conclusion

In the current market structure, retail brokers set up relationships with market makers, and send individual orders to individual market makers. While market makers are evaluated on the aggregate execution quality they deliver, there is no pre-trade communication over individual orders. The SEC concept for order-by-order auctions would require each individual order to be exposed in a bidding process.

Our model shows that a switch to order-by-order auctions comes with trade-offs. Allocative efficiency is improved, as order-by-order auctions ensure that an incoming retail market order is always routed to the market maker who has observed the lowest cost signal. Given the common-value nature of the auction, however, the winner's curse issue is more severe under order-by-order auctions, and market makers obtain higher profits in the auction relative to the broker's routing system. Retail investors can be worse off in the switch to order-by-order auctions, particularly in illiquid stocks or at times when interest in voluntary liquidity provision is low, as market participants could earn more rent in these cases.

We also explore the role of institutional traders in order-by-order auctions and its impact on welfare outcomes and equilibrium structure. The level of information asymmetry between institutional traders and wholesalers plays a key role in determining these outcomes. In the absence of information asymmetry, switching to order-by-order auctions improves total welfare. However, if information asymmetry is high, this result may be reversed. We also examine the cross-subsidization effect among heterogeneous stocks, which implies that switching to order-by-order auctions may affect retail investors with different holdings differently.

We empirically evaluate Retail Liquidity Programs (RLPs) to gain insight into potential bidding participation in the proposed order-by-order auctions. Much like the proposed order-by-order auctions, existing RLPs allow any market participant to bid potential price improvement to incoming retail market orders. While these RLPs offer potential price improving liquidity, this liquidity is very rarely offered in less liquid stocks, and disappears in times of volatility. Analysis of the IEX RLP suggests far fewer bidders have an interest in trading with retail than the SEC's CAT analysis of mid-quote liquidity. While auctions can improve retail investor when bidding participation is high, the current system of broker's routing delivers superior retail investor welfare when bidding participation is low, as it encourages greater competition at the cost of lower allocative efficiency.

REFERENCES

Adams, William James, and Janet L Yellen, 1976, Commodity Bundling and the Burden of Monopoly, *Quarterly Journal of Economics* 90, 475–498.

Aramian, Fatemeh, and Carole Comerton-Forde, 2023, Retail Trading in European Equity Markets, Working Paper.

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, R., and H Jennings, 2023, On the Potential Cost of Mandating Qualified Auctions for Marketable Retail Orders, Available at SSRN: 4403047.

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

Battalio, Robert, and Robert Jennings, 2022, Why do Brokers who do not Charge Payment for Order Flow Route Marketable Orders to Wholesalers?, Technical report, Working Paper.

Battalio, Robert, Robert Jennings, Mehmet Saglam, and Jun Wu, 2022, Identifying Market Maker Trades as "Retail" From TAQ: No Shortage of False Negatives and False Positives.

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

Bernhardt, Dan, and Eric Hughson, 1997, Splitting Orders, *Review of Financial Studies* 10, 69–101.

Biais, Bruno, David Martimort, and Jean-Charles Rochet, 2000, Competing Mechanisms in a Common Value Environment, *Econometrica* 68, 799–837.

Bishop, Allison, 2022, "The SEC Isn't Mad at PFOF. They're Just Disappointed", Proof Trading, [Accessed: 2022 02 01].

Bryzgalova, Svetlana, Anna Pavlova, and Taisiya Sikorskaya, 2022, Retail Trading in Options and the Rise of the Big Three Wholesalers, *Journal of Finance*, Forthcoming.

Corwin, Shane A, and Jay F Coughenour, 2008, Limited Attention and the Allocation of Effort in Securities Trading, *Journal of Finance* 63, 3031–3067.

De Vries, Sven, and Rakesh V Vohra, 2003, Combinatorial Auctions: A Survey, *INFORMS Journal on Computing* 15, 284–309.

Dyhrberg, Anne Haubo, Andriy Shkilko, and Ingrid M Werner, 2022, The Retail Execution Quality Landscape, Fisher College of Business Working Paper.

Easley, David, Nicholas M Kiefer, and Maureen O'Hara, 1996, Cream-skimming or Profit-Sharing? The Curious Role of Purchased Order Flow, *Journal of Finance* 51, 811–833.

Eaton, Gregory W, T Clifton Green, Brian S Roseman, and Yanbin Wu, 2022, Retail Trader Sophistication and Stock Market Quality: Evidence from Brokerage Outages, *Journal of Financial Economics* 146, 502–528.

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

Foley, Sean, Anqi Liu, Katya Malinova, Andreas Park, and Andriy Shkilko, 2020, Cross-Subsidizing Liquidity, Technical report, Working Paper, Macquarie University.

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

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

Glosten, Lawrence R, and Paul R Milgrom, 1985, Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders, *Journal of Financial Economics* 14, 71–100.

Harris, Milton, and Robert M Townsend, 1981, Resource Allocation Under Asymmetric Information, *Econometrica* 49, 33–64.

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

Hendershott, Terrence, Dmitry Livdan, and Norman Schürhoff, 2021, All-to-all Liquidity in Corporate Bonds.

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, *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.

Jehiel, Philippe, Moritz Meyer-Ter-Vehn, and Benny Moldovanu, 2007, Mixed Bundling Auctions, *Journal of Economic Theory* 134, 494–512.

Klemperer, Paul, 1999, Auction Theory: A Guide to the Literature, *Journal of Economic Surveys* 13, 227–286.

Kyle, Albert S, 1985, Continuous Auctions and Insider Trading, *Econometrica* 53, 1315–1335.

McAfee, R Preston, John McMillan, and Michael D Whinston, 1989, Multiproduct Monopoly, Commodity Bundling, and Correlation of values, *Quarterly Journal of Economics* 104, 371–383.

Menezes, Flavio M, and Paulo K Monteiro, 2004, *An Introduction to Auction Theory* (OUP Oxford).

Menkveld, Albert J, Bart Zhou Yueshen, and Haoxiang Zhu, 2017, Shades of Darkness: A Pecking Order of Trading Venues, *Journal of Financial Economics* 124, 503–534.

Myerson, Roger B, 1979, Incentive Compatibility and the Bargaining Problem, *Econometrica* 47, 61–73.

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

Pekeč, Aleksandar, and Michael H Rothkopf, 2003, Combinatorial Auction Design, *Management Science* 49, 1485–1503.

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

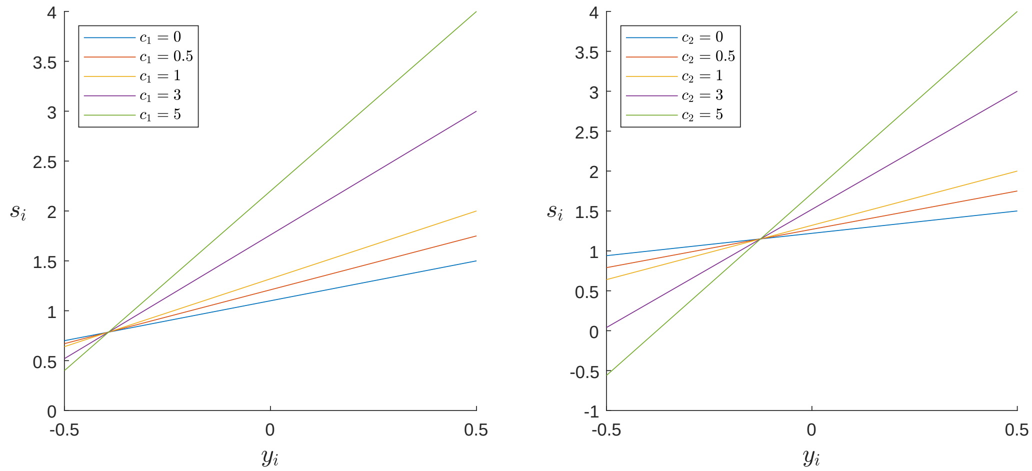
Schwenk-Nebbe, Sander, 2021, The Participant Timestamp: Get The Most Out Of TAQ Data, Available at SSRN 3984827.

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.

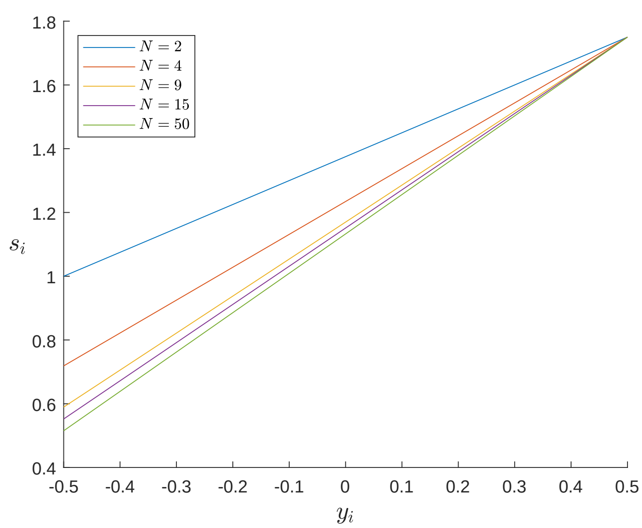
SIFMA, 2022, "SIFMA Equity Market Structure Roundtable", Washington, D.C. [Accessed: 2023 12 01].

Securities and Exchange Commission, 2022, Order Competition Rule, Proposal for Rule 615. [Accessed: 2022 12 16].

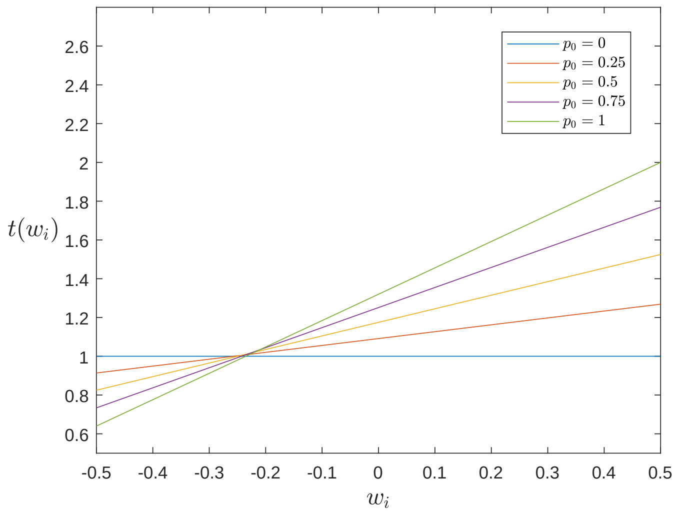
Appendix A. Graphs



Plot 1. Equilibrium Strategy $s_i(y_i)$ under Order-by-order Auctions with Different c_1 and c_2. We use $c_0 = 1$ and $N = 5$ for both panels. For the left panel, we fix $c_2 = 1$ and for the right panel, we fix $c_1 = 1$.



Plot 2. Equilibrium Strategy $s_i(y_i)$ under Order-by-order Auctions with Different N. We use $c_0 = 1$, $c_1 = 0.5$ and $c_2 = 1$ for both panels.



Plot 3. $t(w_i)$ for Different Values of p_0 under Broker's Routing. We use $c_0 = 1$, $c_1 = 1$, $c_2 = 1$ and $N = 5$ for this plot.

Appendix B. Proofs

Proof of Proposition 1

Consider any $i \in \{1, 2...N\}$ and $(x, y) \in \left[-\frac{1}{2}, \frac{1}{2}\right] \times \left[-\frac{1}{2}, \frac{1}{2}\right]$, let $G(x|y) = \text{Prob}\left[\min_{-i} y_{-i} \leq x | y_i = y\right]$ and $g(x|y) = \frac{dG(x|y)}{dx}$. It is easy to show $G(x|y) = 1 - \left(\frac{1}{2} - x\right)^{N-1}$, $g(x|y) = (N-1)\left(\frac{1}{2} - x\right)^{N-2}$. Let

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

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

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

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

$$-\beta(y) + \left(\frac{1-G(y|y)}{g(y|y)}\right)\beta'(y) + v(y|y) = 0. \tag{B1}$$

We first conjecture that $\beta(y)$ is linear, i.e., there exist k_0, k_1 such that $\beta(y) = k_0 + k_1 y$. Substitute this into (B1), we have

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

Then k_0 and k_1 are solved by $-k_0 + \frac{\frac{1}{2}k_1}{N-1} + c_0 + c_1\frac{N-2}{4N} = 0$ and $-k_1 - \frac{k_1}{N-1} + \frac{c_1}{2} + \frac{c_1}{N} + c_2 = 0$. Then we get

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

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

It's easy to check that $\left.\frac{\partial U_i(z,y)}{\partial z}\right|_{z=y} = 0$ is also the sufficient condition in the optimization problem in this linear equilibrium because of the linearity of the equilibrium.

Proof of Lemma 1

First we 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 expected total welfare of the investor is

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

The total expected welfare of market makers is

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

and the total welfare 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 3

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

$$\begin{aligned}
&W_I^{BR} < W_I^{OBO} \\
\Longleftrightarrow & -\left[c_0 + p_0 \frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)}\right] < -\left[c_0 + \frac{1}{N(N+1)}c_1 - \frac{N-3}{2(N+1)}c_2\right] \\
\Longleftrightarrow & N(N-3)\frac{c_2}{c_1} > 2(1-p_0).
\end{aligned}$$

Proof of Proposition 5

When $\delta_c = 0$, the institutional traders and wholesalers receive i.i.d signals, and they are symmetric. We first conjecture that all market makers choose the same linear equilibrium strategy

$$\tilde{\beta}_i\left(y_i; \delta_c = 0\right) = \tilde{k}_0\left(\delta_c = 0\right) + \tilde{k}_1\left(\delta_c = 0\right) y_i.$$

The number of market makers is $N + N_0$. We follow the proof of Proposition 1, the function $v\left(x, y\right)$ now becomes

$$\begin{aligned}
\tilde{v}\left(x,y\right) &= \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} y_{-i} = x, y_i = y\right] \\
&= c_0 + c_1 \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N+N_0} y_j | \min_{-i} y_{-i} = x, y_i = y\right] + c_2 \mathbb{E}\left[y_i | \min_{-i} y_{-i} = x, y_i = y\right] \\
&= c_0 + \left(\frac{c_1}{N+N_0} + c_2\right) y + \frac{c_1}{N+N_0} x + c_1 \frac{N+N_0-2}{N+N_0}\frac{1}{2}\left(\frac{1}{2} + x\right) \\
&= \left(c_0 + c_1 \frac{N+N_0-2}{4\left(N+N_0\right)}\right) + \frac{c_1}{2} x + \left(\frac{c_1}{N+N_0} + c_2\right) y,
\end{aligned}$$

which is the $v\left(x,y\right)$ function with $\left(N+N_0\right)$ wholesalers. For the rest of the proof, we follow the proof of Proposition 1, and we can show that the equilibrium strategy is equivalent to that in Proposition 1 with the number of wholesalers being $N + N_0$.

Proof of Proposition 6

Since the equilibrium of broker's routing is the same as that in the baseline model, we have

$$\tilde{W}^{BR}_{total} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right),$$

$$\tilde{W}^{BR}_{I} = -\left[c_0 + p_0 \frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)}\right],$$

and

$$\tilde{W}^{BR}_{W} = \frac{p_0\left(2c_1 - p_0 c_1 + Nc_2\right)}{N\left(1+N\right)}.$$

For order-by-order auctions, the welfare outcomes are

$$\tilde{W}^{OBO}_{total} = -\left(c_0 - \frac{N+N_0-1}{N+N_0+1}\frac{c_2}{2}\right),$$

$$\tilde{W}^{OBO}_{I} = -\left[c_0 + \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)}c_1 - \frac{N+N_0-3}{2\left(N+N_0+1\right)}c_2\right],$$

and

$$\tilde{W}^{OBO}_{W} = \frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{c_1}{N+N_0} + c_2\right).$$

First, it's obvious that

$$\tilde{W}^{OBO}_{total} > \tilde{W}^{BR}_{total},$$

because $p_0 \in (0,1)$ and $N_0 > 1$. Second,

$$\begin{aligned}
&\tilde{W}^{BR}_{W} < \tilde{W}^{OBO}_{W} \\
\Longleftrightarrow &\frac{p_0\left(2c_1 - p_0 c_1 + Nc_2\right)}{N\left(1+N\right)} < \frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{c_1}{N+N_0} + c_2\right) \\
\Longleftrightarrow &\frac{p_0\left(2 - p_0 + N\frac{c_2}{c_1}\right)}{N\left(1+N\right)} < \frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{1}{N+N_0} + \frac{c_2}{c_1}\right) \\
\Longleftrightarrow &\frac{p_0\left(2-p_0\right)}{N\left(1+N\right)} + \frac{p_0}{1+N}\frac{c_2}{c_1} < \frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{1}{N+N_0} + \frac{c_2}{c_1}\right) \\
\Longleftrightarrow &\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}\right)\frac{c_2}{c_1} > -\left(\frac{N}{\left(N+N_0\right)^2}\frac{1}{N+N_0+1} - \frac{p_0\left(2-p_0\right)}{N\left(1+N\right)}\right) \\
\Longleftrightarrow &\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}\right)\frac{c_2}{c_1} > -\frac{1}{N+N_0}\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{\left(1+N\right)}\frac{\left(2-p_0\right)\left(N+N_0\right)}{N}\right)
\end{aligned}$$

Since

$$\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}>\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N},$$

we know that

$$\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}\right)\frac{c_2}{c_1}>-\frac{1}{N+N_0}\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)$$

is equivalent to

$$\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}>0\Longleftrightarrow p_0<\frac{N}{N+N_0}\frac{1+N}{N+N_0+1}$$

and

$$\frac{c_2}{c_1}>\frac{-\frac{1}{N+N_0}\left(\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)}{\frac{N}{N+N_0}\frac{1}{N+N_0+1}-\frac{p_0}{1+N}}.$$

Finally,

$$\begin{aligned}
&\tilde{W}_I^{BR}<\tilde{W}_I^{OBO}\\
\Longleftrightarrow&-\left[c_0+p_0\frac{2(2-p_0)c_1-(N-3)Nc_2}{2N(1+N)}\right]<-\left[c_0+\frac{1}{(N+N_0)(N+N_0+1)}c_1-\frac{N+N_0-3}{2(N+N_0+1)}c_2\right]\\
\Longleftrightarrow&p_0\frac{2(2-p_0)c_1-(N-3)Nc_2}{2N(1+N)}>\frac{1}{(N+N_0)(N+N_0+1)}c_1-\frac{N+N_0-3}{2(N+N_0+1)}c_2\\
\Longleftrightarrow&p_0\frac{2(2-p_0)-(N-3)N\frac{c_2}{c_1}}{2N(1+N)}>\frac{1}{(N+N_0)(N+N_0+1)}-\frac{N+N_0-3}{2(N+N_0+1)}\frac{c_2}{c_1}\\
\Longleftrightarrow&\left(\frac{N+N_0-3}{2(N+N_0+1)}-p_0\frac{(N-3)}{2(1+N)}\right)\frac{c_2}{c_1}>\frac{1}{(N+N_0)(N+N_0+1)}-p_0\frac{(2-p_0)}{N(1+N)}\\
\Longleftrightarrow&\frac{c_2}{c_1}>\frac{\frac{1}{(N+N_0)(N+N_0+1)}-\frac{p_0(2-p_0)}{N(1+N)}}{\frac{N+N_0-3}{2(N+N_0+1)}-p_0\frac{(N-3)}{2(1+N)}}.
\end{aligned}$$

The last inequality holds because we always have $\frac{N+N_0-3}{2(N+N_0+1)}-p_0\frac{(N-3)}{2(1+N)}>0$.

Proof of Proposition 7

We need to verify that this is indeed an equilibrium. Let $\underline{\delta} = \max\{\delta_{c1}, \delta_{c2}\}$ where δ_{c1} is defined by (B3) and δ_{c2} is defined by (B4).

Let δ_{c1} be the value that satisfies

$$k_0^- + k_1^- \cdot \frac{1}{2} = k_0^+ - k_1^+ \cdot \frac{1}{2},$$

i.e.,

$$c_0 - \delta_{c1} + \frac{c_1}{4N_0}\left(N_0 - 1 + \frac{2}{N_0}\right) + \frac{c_2}{2N_0} + \frac{N_0 - 1}{N_0}\left(\frac{c_1}{2}\frac{N_0 + 2}{N_0} + c_2\right)\frac{1}{2} \tag{B2}$$

$$= c_0 + \delta_{c1} + \frac{c_1}{4(N + N_0)}\left(N + N_0 - 1 + \frac{2}{N + N_0}\right) + \frac{c_2}{2(N + N_0)} - \frac{N + N_0 - 1}{N + N_0}\left(\frac{c_1}{2}\frac{N + N_0 + 2}{N + N_0} + c_2\right)\frac{1}{2}. \tag{B3}$$

Then when $\delta_c > \delta_{c1}$, $\left[k_0^- - k_1^- \cdot \frac{1}{2}, k_0^- + k_1^- \cdot \frac{1}{2}\right] \cup \left[k_0^+ - k_1^+ \cdot \frac{1}{2}, k_0^+ + k_1^+ \cdot \frac{1}{2}\right] = \emptyset$. This implies that the under the equilibrium conjectured, when $\tilde{c}_0 = c_0 - \delta_c$, only institutional traders can obtain the order no matter what signals market participants observe.

We first verify that it is optimal for any institutional trader to choose $\tilde{s}^-(y; \delta_c)$ if observing $c_0 + \delta_c$, given other market participants' strategies. When $\tilde{c}_0 = c_0 - \delta_c$, then clearly $\tilde{s}^-(y; \delta_c)$ is an equilibrium if we only have N_0 institutional traders in the market, as suggested by Proposition 1. This is essentially the baseline model of order-by-order auctions with N_0 bidders and unconditional expected inventory cost being $c_0 - \delta_c$. This means that it's optimal for any institutional trader to choose $\tilde{s}^-(y; \delta_c)$ if there are only $N_0 - 1$ other institutional traders who also choose $\tilde{s}^-(y; \delta_c)$ and no wholesalers in the market. Adding N wholesalers choosing $\tilde{s}^+(y; \delta_c)$ does not change this optimality, because given other institutional traders' choice $\tilde{s}^-(y; \delta_c)$, the N wholesalers will never obtain any order in any state when $\tilde{c}_0 = c_0 - \delta_c$.

Then we verify that it's optimal for any institutional trader to choose $\tilde{s}^+(y; \delta_c)$ if observing $c_0 + \delta_c$, given other market participants' strategies. Following our Proposition 1 in the baseline model of order-by-order auctions, $\tilde{s}^+(y; \delta_c)$ is an equilibrium with $N + N_0$ market makers and unconditional expected inventory cost being $c_0 + \delta_c$. So it's optimal for any institutional trader to choose $\tilde{s}^+(y; \delta_c)$ if there are other $N + N_0 - 1$ market makers also choosing $\tilde{s}^+(y; \delta_c)$.

Finally, we verify that it's optimal for any wholesaler i to choose $\tilde{s}^+(y; \delta_c)$, given other market participants' strategies. Suppose the wholesaler i observes a signal y_i, then the wholesaler's utility is $U_i = \frac{1}{2}U_1(s_i) + \frac{1}{2}U_2(s_i)$, where U_1 (U_2) is wholesaler i's profit when the state is $c_0 - \delta_c$ ($c_0 + \delta_c$). It's clear that for any y_i, we have $\tilde{s}^+(y_i;\delta_c) = \arg\max_{s_i \in \left(k_0^- + k_1^- \cdot \frac{1}{2}, \infty\right)} U_i$. This is because when $s_i \in \left(k_0^- + k_1^- \cdot \frac{1}{2}, \infty\right)$, $\left[k_0^- - k_1^- \cdot \frac{1}{2}, k_0^- + k_1^- \cdot \frac{1}{2}\right] \cup \left[k_0^+ - k_1^+ \cdot \frac{1}{2}, k_0^+ + k_1^+ \cdot \frac{1}{2}\right] = \emptyset$, and thus we always have $U_1(s_i) = 0$. And by definition $\tilde{s}^+(y_i;\delta_c) = \arg\max_{s_i \in \left(k_0^- + k_1^- \cdot \frac{1}{2}, \infty\right)} U_2$. It is also clear that the wholesaler will never choose $s_i < k_0^- - k_1^- \cdot \frac{1}{2}$, as any $s_i < k_0^- - k_1^- \cdot \frac{1}{2}$ is dominated by $s_i = k_0^- - k_1^- \cdot \frac{1}{2}$. Suppose that the wholesaler chooses $s_i \in \left[k_0^- - k_1^- \cdot \frac{1}{2}, k_0^- + k_1^- \cdot \frac{1}{2}\right]$. Note that $k_0^- + k_1^- \cdot \frac{1}{2} - \left[(c_0 - \delta_c) - \frac{c_1 + c_2}{2}\right] > 0$, and the upper bound of the profit in the case $c_0 - \delta_c$ is $k_0^- + k_1^- \cdot \frac{1}{2} - \left[(c_0 - \delta_c) - \frac{c_1+c_2}{2}\right]$ because $k_0^- + k_1^- \cdot \frac{1}{2}$ is the highest spread in $s_i \in \left[k_0^- - k_1^- \cdot \frac{1}{2}, k_0^- + k_1^- \cdot \frac{1}{2}\right]$ and $(c_0 - \delta_c) - \frac{c_1+c_2}{2}$ is the lowest inventory cost. Besides, in the case $c_0 + \delta_c$, the wholesaler i will obtain the order with probability one. And one upper bound of the profit is $k_0^- + k_1^- \cdot \frac{1}{2} - \left[(c_0 + \delta_c) - \frac{c_1+c_2}{2}\right]$. Then $U_1 \leq k_0^- + k_1^- \cdot \frac{1}{2} - \left[(c_0 - \delta_c) - \frac{c_1+c_2}{2}\right]$ and $U_2 \leq k_0^- + k_1^- \cdot \frac{1}{2} - \left[(c_0 + \delta_c) - \frac{c_1+c_2}{2}\right]$. This implies $U_i \leq \frac{1}{2}U_1 + \frac{1}{2}U_2 \leq k_0^- + k_1^- \cdot \frac{1}{2} - \left[c_0 - \frac{c_1+c_2}{2}\right]$. Then

$$U_i < 0 \Longleftarrow k_0^- + k_1^- \cdot \frac{1}{2} - \left[c_0 - \frac{c_1 + c_2}{2}\right] < 0 \Longleftrightarrow \delta_c < \delta_{c2},$$

where

$$\delta_{c2} = \frac{c_1}{4N_0}\left(N_0 - 1 + \frac{2}{N_0}\right) + \frac{c_2}{2N_0} + k_1^- \cdot \frac{1}{2} + \frac{c_1 + c_2}{2}. \tag{B4}$$

Then when $\delta_c > \underline{\delta} = \max\{\delta_{c1}, \delta_{c2}\}$, we have $\tilde{s}^+(y_i;\delta_c) = k_0^+ + k_1^+ y_i = \arg\max_{s_i \in (-\infty,\infty)} U_i$, which implies that it is optimal for any wholesaler i to choose $\tilde{s}^+(y;\delta_c)$, given other market participants' strategies.

Proof of Lemma 4

We introduce the random variable $r = \min_i w_i \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ and its CDF $H(r) = 1 - \left(\frac{1}{2} - r\right)^N$ and PDF $h(r) = N\left(\frac{1}{2} - r\right)^{N-1}$. First, we know that in our baseline model of broker's routing, the total welfare is

$$W_{total}^{BR} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

The total welfare only depends on inventory allocation but not equilibrium spread, as the equilibrium spread is just a transfer between market makers and investors. In our extension of heterogeneous stocks, it is still the market maker with lowest liquidity signal realization y that obtains the order, so the order allocation is the same as that in our baseline model for any stocks (c_0, c_1, c_2). Then the total welfare is this extension satisfies

$$W^{BR}_{heter,total} = W^{BR}_{total} = -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right).$$

Since the equilibrium bidding strategy is $T(w) = \bar{K}_0 + \bar{K}_1 w$, the investor's welfare is

$$\begin{aligned} W^{BR}_{heter,I} &= -\mathbb{E}\left[\bar{K}_0 + \bar{K}_1 r | \min_i w_i = r\right] \\ &= -\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\bar{K}_0 + \bar{K}_1 r\right) N \left(\frac{1}{2} - r\right)^{N-1} dr \\ &= -\left[\bar{c}_0 + p_0 \frac{2(2-p_0)\bar{c}_1 - (N-3)N\bar{c}_2}{2N(1+N)}\right]. \end{aligned}$$

By $W^{BR}_{heter,M} = W^{BR}_{heter,total} - W^{BR}_{heter,I}$, we know

$$\begin{aligned} W^{BR}_{heter,M} &= -\left(c_0 - p_0 \frac{N-1}{N+1}\frac{c_2}{2}\right) + \left[\bar{c}_0 + p_0 \frac{2(2-p_0)\bar{c}_1 - (N-3)N\bar{c}_2}{2N(1+N)}\right] \\ &= (\bar{c}_0 - c_0) + \frac{p_0}{2(N+1)}\left[(N-1)c_2 - (N-3)\bar{c}_2\right] + p_0 \frac{(2-p_0)\bar{c}_1}{N(1+N)}. \end{aligned}$$

Internet Appendix: Would Order-By-Order Auctions Be Competitive?

Thomas Ernst, Chester Spatt and Jian Sun

Appendix I.A.1. RLP Details

NYSE was the first to operate a RLP, on August 1, 2012.[25] 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 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

[25] 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.

requirement to provide liquidity for a certain percentage of the trading day, and are therefore potentially more accessible to non-market-making firms. In approving the NASDAQ RLP, the SEC notes that "the Program might also create a desirable opportunity for institutional investors to interact with retail order flow that they are not able to reach currently. Today, institutional investors often do not have the chance to interact with marketable retail orders that are executed pursuant to internalization arrangements. Thus, by submitting RPI Orders, institutional investors may be able to reduce their possible adverse selection costs by interacting with retail order flow" SEC (2013). The SEC identifies the same desirable feature, that of more potential counter parties for retail trades, that are highlighted in a potential move to order-by-order competition.

The Investors Exchange (IEX) offers a retail liquidity program whereby retail liquidity providers can enter hidden mid-point peg limit orders which are only available to retail market orders. All mid-point peg orders enter the same time priority queue, whether or not they are only available to retail investors, and both have queue priority over the IEX D-limit order, which is the discretionary limit order whereby IEX has a proprietary algorithm to modify the price of these orders whenever it forecasts a potentially disadvantageous change in quotes. The IEX RLP only takes mid-point orders, and disseminates a RLP indicative flag when there is at least one round lot of RLP interest. All eligible retail orders have no trading fees, either for the retail broker or the retail liquidity provider.

The IEX RLP is the most recent program, first offering the RLP trading functionality on October 1, 2019. IEX initially had no RLP indicators, but added indicators on October 13, 2021. Unlike other retail liquidity programs, the IEX program only allows mid-quote prices. Therefore, while the size available is hidden, an advertised RLP indicator from IEX confirms that at least 100 shares are available at the specific price of the mid-quote. To offer RLP indicators, the program required an approved exemption from SEC Rule 242.602, as the RPI would indicate a specific price and a minimum quantity of shares, but would not be accessible to non-retail marketable orders.

The Members Exchange MEMX applied to create an RLP program, but was denied by the SEC on February 14, 2022. The MEMX proposal differed from previous proposals in the determination of price-time priority. Under the MEMX proposal, incoming retail market orders first interact with hidden RPI orders before interacting with hidden non-RPI orders, even if the hidden non-RPI are at the same price level and have time priority.

MEMX argued that because hidden RPI orders do contribute to the dissemination of the RPI interest indicator, they should have priority over hidden non-RPI orders at each price level, analogous to standard practice of non-hidden orders having priority over hidden orders at each price level. The SEC disagreed, and ruled that the change in priority would violate Section (6)(b)(5) and Section 11A of the Exchange Act.[26]

Table III presents the exact total volumes in our sample executed when RPI programs are active, and when they are not.[27]

Retail Liquidity Programs have indicative interest for a large portion of the trading day. Figure 5 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. Some of the differences in RPI shares across assets may come from the different rules. There is also a considerable listing-exchange advantage. NYSE Arca's retail liquidity program, for example, has RPI interest for less than 20% of the trading day for Russell 1000 or Russell 2000 stocks, but has RPI interest for over 75% of the trading day for ETFs in our sample.

There is a considerable discrepancy between the share of time that retail liquidity programs have RPL flags active, and the share of trading volume which executes in RLP. Figure 6 depicts the trading volume split of trades when the exchange's RPI Flag is active, and the trading volume split when it is not. Sub-penny executions are less than 1% of total trading volume at exchanges, even when the RPI Flag is active.[28] On-exchange mid-quote trading

[26]Ironically, in the Order-by-Order proposal from the SEC, auctions must give auction responses higher priority than hidden limit orders (Securities and Exchange Commission (2022), Proposed Rule 615 Section IV C.5.) In other words, MEMX's RLP application was denied because it proposed giving resting RPI orders priority over hidden resting limit orders, but auctions would be *required* to give auction responses priority over hidden resting limit orders.

[27]To exclude biases from incorrect interpretation of 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 is reported as 1 share.

[28]Note that exchange sub-penny volume when there is no RPI Flag is small but non-zero, which arise when there is an odd-lot limit order in the RLP program.When there is odd-lot limit order interest, the RLP flag will not be disseminated (as there are less than 100 shares of interest) but retail market orders can still execute against this odd-lot volume. Another possible explanation for this discrepancy is inaccuracy in the timestamp-based matching of the sort described by Schwenk-Nebbe (2021), who show that the exchange processing and dissemination of quotes is typically several microseconds faster than that of trades.

Table III: Summary Volumes By Each Price Increment. This table presents summary total trading volume (in billions of dollars) in our sample for each sub-penny category of trade: at-quote, mid-quote, and sub-penny. 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.

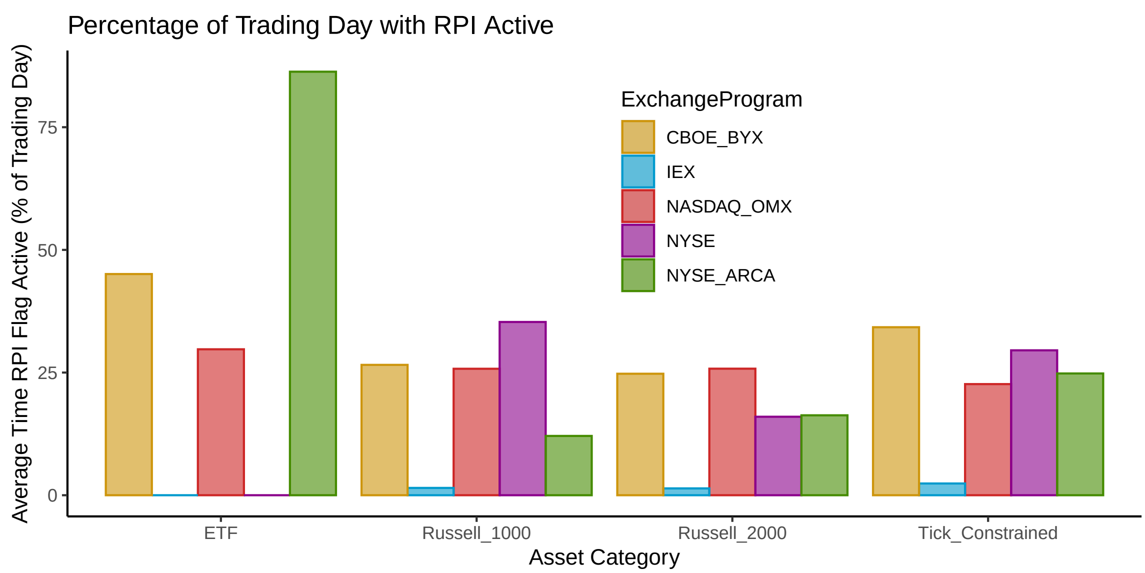
Panel A: Exchange Trades. We define a flag as 'active' if there is contemporaneous RLP interest at the exchange where trade occurs.

Asset	RLP	Volume			Percent		
	Flag	mid-quote	At Quote	Subpenny	mid-quote	At Quote	Subpenny
ETF	Active	2259	35868	414	3.2	51	0.6
ETF	None	317	9160	55	0.5	13	0.1
Russell_1000	Active	3162	60890	235	1.7	32	0.1
Russell_1000	None	1933	50111	99	1.0	27	0.1
Russell_2000	Active	283	5376	13	1.5	28	0.1
Russell_2000	None	306	6012	2.6	1.6	31	0.01
TickConstr.	Active	2767	40510	400	3.3	48	0.5
TickConstr.	None	676	12773	67	0.8	15	0.1

Panel B: Off-Exchange Trades. We define a flag as 'active' if there is contemporaneous RLP interest disseminated by any exchange.

Asset	RLP	Volume			Percent		
	Flag	mid-quote	At Quote	Subpenny	Mid-quote	At Quote	Subpenny
ETF	Active	3626	9070	5751	5.2	13	8.2
ETF	None	537	1909	1134	0.8	2.7	1.6
Russell_1000	Active	8638	20795	10127	4.6	11	5.4
Russell_1000	None	5895	16953	8756	3.1	9.0	4.7
Russell_2000	Active	743	2161	822	3.8	11	4.2
Russell_2000	None	723	2107	833	3.7	11	4.3
TickConstr.	Active	4740	9643	6456	5.6	12	7.7
TickConstr.	None	1257	3079	1869	1.5	3.7	2.2

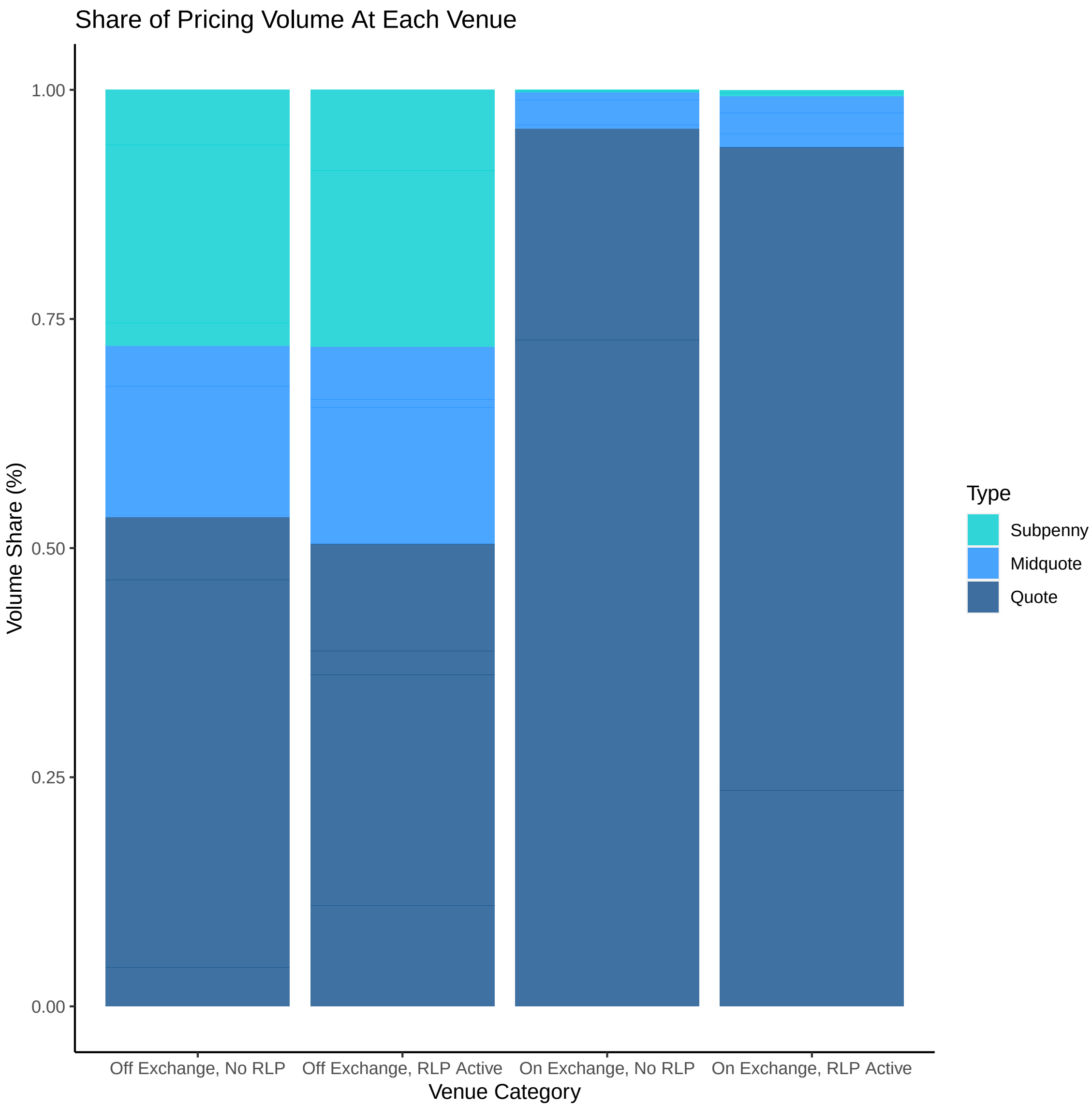
Figure 5. Time Share of Retail Liquidity Programs. We plot the average time that an RPI indicator is active, measured as percentage of time active out of the total trading day. Our sample can be divided into three groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the day for at least one-third of the days of our sample.



volume is considerably higher when the RPI Flag is active, but still represents less than 2% of total trading volume. Furthermore, this mid-quote volume is a mixture of both retail interest, including from the IEX RLP which only allows retail RPI orders to be priced at mid-quote, and non-RLP program hidden mid-quote liquidity. The vast majority of mid-quote and sub-penny trading occurs off-exchange. When RPI Flags are active, a larger share of off-exchange volume occurs at sub-penny or mid-quote prices.

The total volume share of Retail Liquidity Programs is stable during our sample period. As Panel A of Figure 7 depicts, on-exchange sub-penny retail trades average below 0.2% of total volume for the Russell 1000 and Russell 2000 stocks in our sample. The volume share of ETFs and tick-constrained stocks is slightly higher, fluctuating between 0.2% to 0.5% of total trading volume. We define a stock as tick-constrained if it has a one penny bid-ask spread for at least 50% of the trading day for at least one-third of the trading days in our sample. For these stocks, competition for a marketable order is potentially larger due to the

Figure 6. Volume Share of Venues. We plot the percentage of volume which executes either at the quote, at the mid-quote, or at a sub-penny price for both on-exchange and off-exchange venues. On both types of venues, a higher percentage of volume occurs at the quote when there is no RPI Flag active, and a higher share of volume executes at sub-penny and mid-quote prices when the RPI Flag is active.



tick constraint, with increased interest in providing liquidity in an RLP. In Panel B of Figure 7, we plot the volume of any exchange sub-penny or mid-quote executions while the RPI

Flag is active. This will include some non-retail hidden liquidity, as well as retail trading 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 RPI Flag is active is around 0.5% to 1.0% of trading volume. While this is a small share of total trading volume, it represents a much larger fraction of retail-only trading volume.

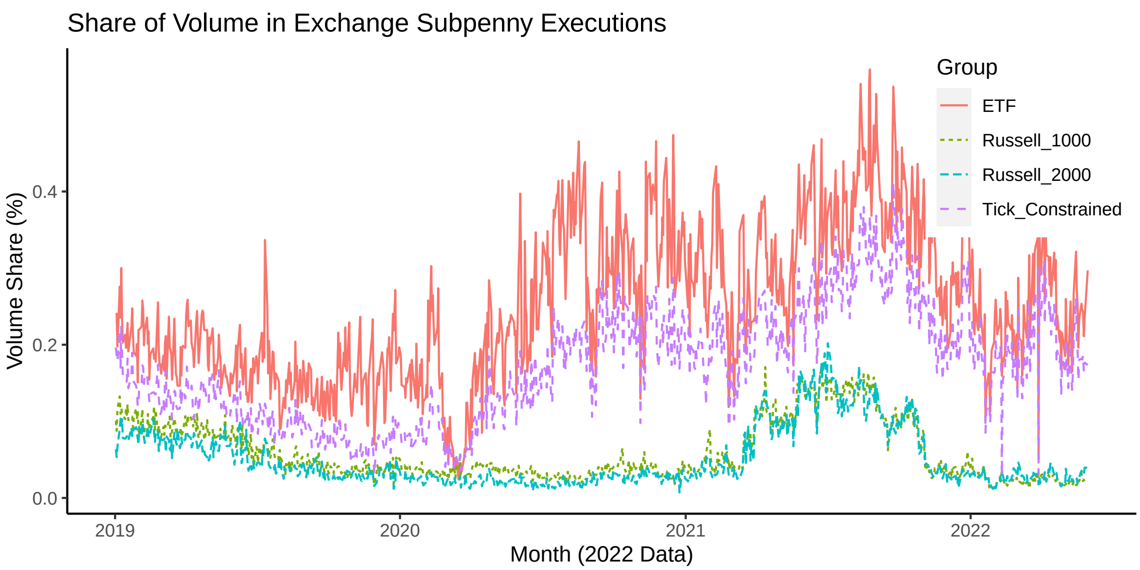
While the Retail Liquidity Programs are the only way that on-exchange trades can be priced in sub-penny increments, retail trades can trade in a variety of methods, with Barber, Huang, Jorion, Odean, and Schwarz (2022) estimating that less than 35% of retail trading takes place at sub-penny prices. Figure 8 depicts the distribution of order sizes for on-exchange and off-exchange sub-penny orders, as a fraction of the NBBO. While a large fraction of sub-penny trades in both venues are odd-lot trades, a far larger share of off-exchange sub-penny trades are for a quantity of shares larger than available at the best bid or offer. Over 2.1% of off-exchange sub-penny trades are for more than 5 times the available shares than the respective national best bid or offer, while only 0.7% of on-exchange sub-penny are for larger than the respective national best bid or offer.

We also investigate the interaction between RLP trading volume and prior or subsequent quoted bid-ask spreads, both when the RPI Flag is active and inactive. Figure 9 presents the ratio of quoted spreads before and after trades. We first divide trading volume into on-exchange and off-exchange trades, and then further divide volume into sub-penny, mid-quote, and at-quote bins. For each individual stock, we observe the quoted bid-ask spread q_{t+i}, where i can be ± 30 seconds, ± 3 milliseconds, or ± 1 milliseconds. We then calculate the average spread $\bar{q}_{t+i}$ separately for when the RPI Flag is or is not active, and plot the ratio $\frac{\bar{q}_i}{\bar{q}}$. When the retail flag is active, off-exchange spreads are very stable, with the same bid-ask spread before and after a trade. When the retail flag is not active, off-exchange spreads before and after a trade tend to be around 2 to 4% wider on average, for all categories of pricing. The large discrepancy in quoted spread ratios before and after a trade provides additional suggestive evidence for the pecking order of Menkveld et al. (2017). The discrepancy in spread ratios around the timing of on-exchange mid-quote and sub-penny trades occur at a momentarily liquid time, when quoted spreads are narrow. In contrast with on-exchange trading, the off-exchange trading spread ratios are far more consistent, with the quoted spread width at the time of trades being very similar to the quoted spread width before or after the trade.

We investigate the distribution of trade prices as a function of the IEX RLP status in the public data. Figure 10 presents the pattern of trade prices as a function of the IEX RLP status. When the IEX RLP has two-sided liquidity (that is, interest in both buying from retail or selling to retail at the mid-quote), trading at mid-quote, whether by retail or non-retail participants, comprises over 50% of all trading volume. There is also no volume at other exchange RLPs, consistent with the IEX's superior price. There are some off-exchange trades at mid-quote when the IEX RLP has two-sided liquidity, though Battalio et al. (2022) document that many sub-penny trades are non-retail.

Figure 7. Volume Share of Sub-penny Retail Liquidity Programs. For each day, we plot the share of volume which executes in a retail liquidity program, out of total volume. Our sample can be divided into three groups: stocks in the Russell 1000 index, stocks in the Russell 2000 index, and ETFs from our sample. We provide a fourth (overlapping) group, "Tick Constrained", comprised of any stock or ETF which meets the criteria of having a quoted bid-ask spread of one penny at least 50% of the day for at least one-third of the days of our sample. Panel A presents the volume share of only exchange sub-penny executions while a RLP indicator is active, while Panel B presents the volume share of all exchange sub-penny or mid-quote executions while a RLP indicator is active.

Panel A: Volume Share for Exchange Sub-penny (Non-Midquote) Executions with an RPI Flag Active



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

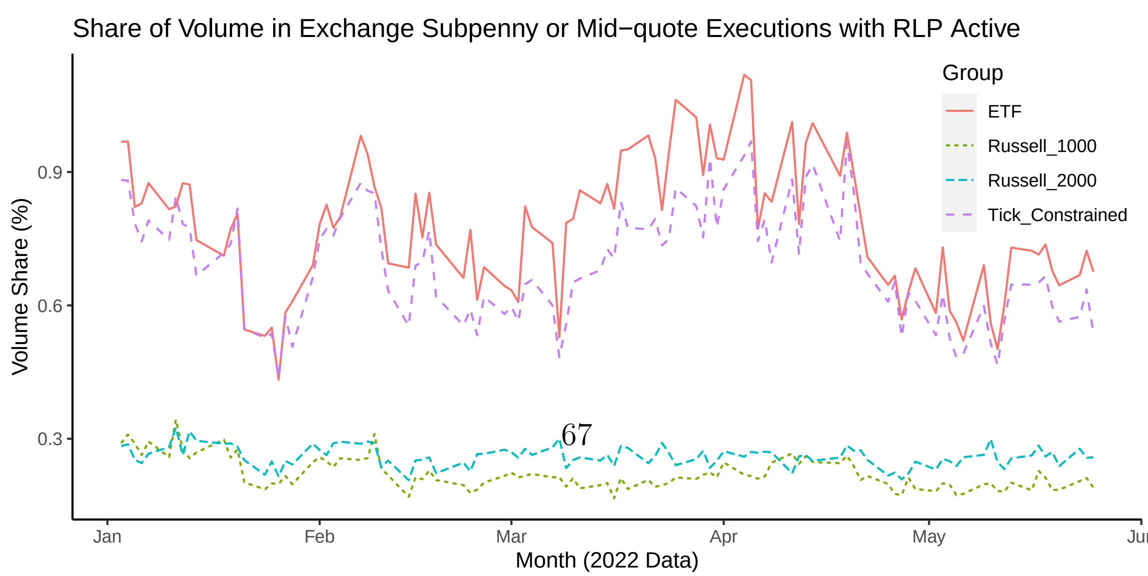


Figure 8. Size Distribution. We plot the distribution of order sizes, as a percentage of the NBBO, for all sub-penny trades occurring in the stocks of our sample on January 3, 2022. We truncate the distribution of orders at 5 times the NBBO. Of all sub-penny trades, 2.1% of all off-exchange sub-penny trades are larger than five times the NBBO, while 0.7% of on-exchange sub-penny trades are larger than five times the NBBO.

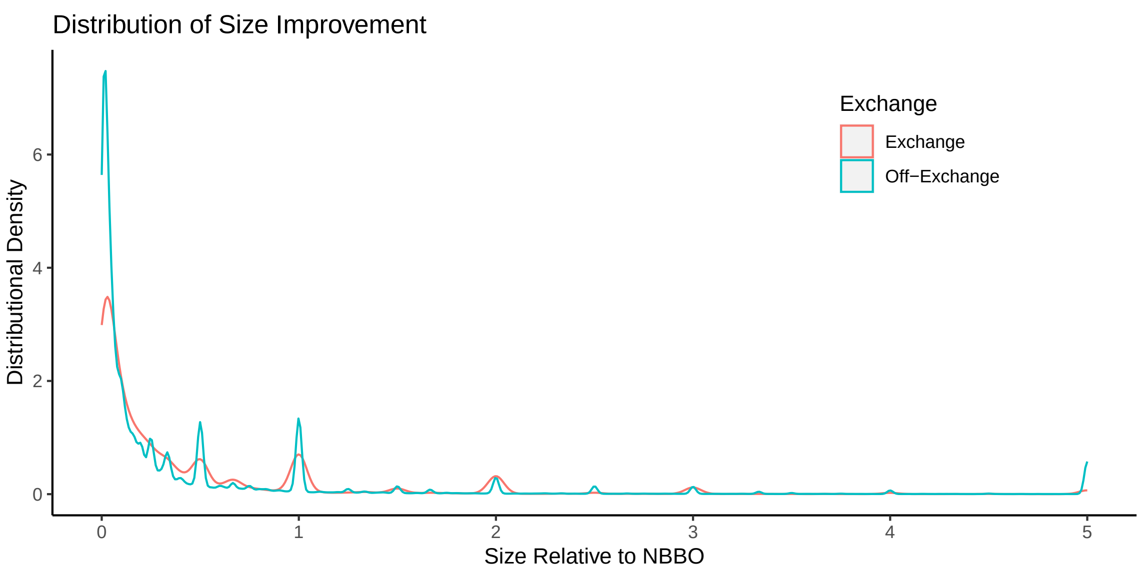
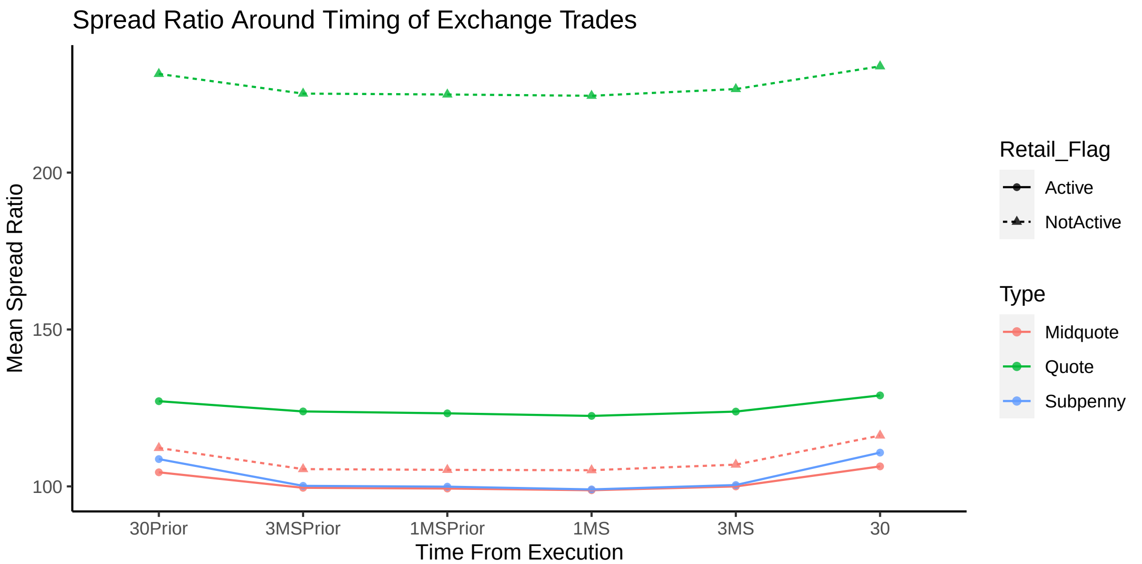


Figure 9. Spread Ratios Around Trades. We plot the change in spreads around the timing of a trade, separately for trades occurring when the RPI Flag is active, and trades occurring when it is not active. For trades occurring at time t, we calculate the quoted spread q_t as well as the quoted spread q_{t+i} occurring at a time-offset of i. We then calculate the mean quoted spreads $\bar{q}$ and $\bar{q}_{+i}$, and plot their ratio $r = \frac{\bar{q}_i}{\bar{q}}$. We consider time offsets of 30 seconds prior to trade, 3 milliseconds prior to trade, 1 millisecond prior to trade, 1 millisecond after trade, 3 milliseconds after trade, and 30 seconds after trade.

Panel A: Spread Ratio For Exchange Trades



Panel B: Spread Ratio For Off-Exchange Trades

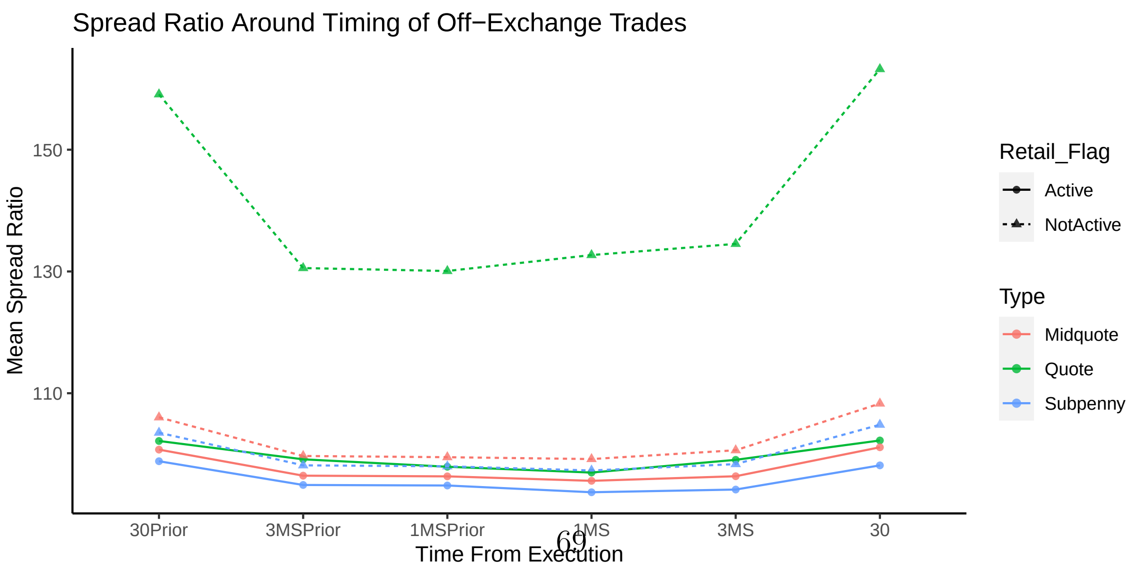
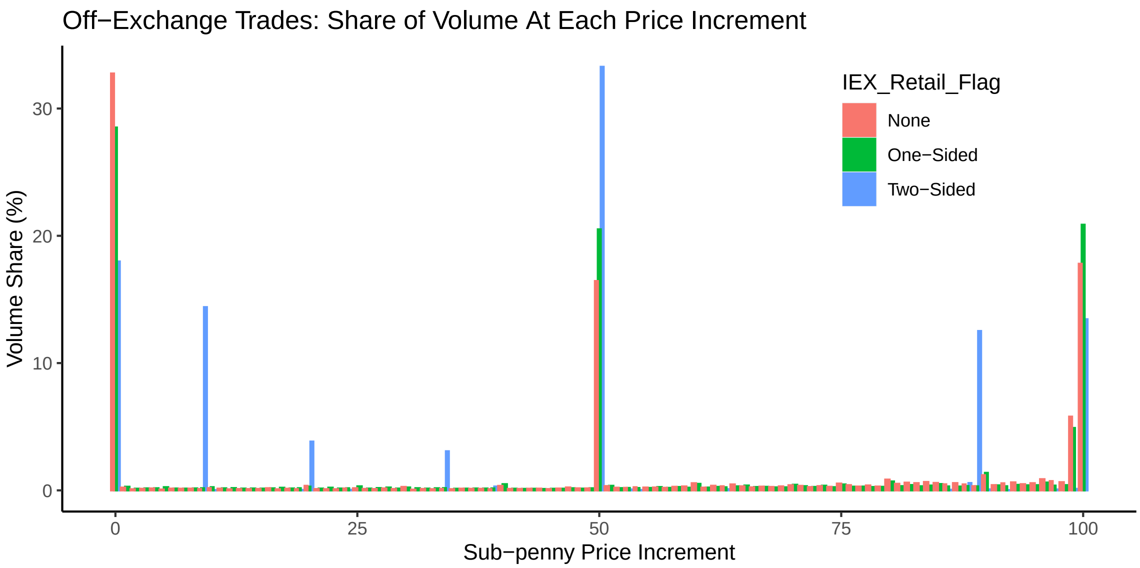
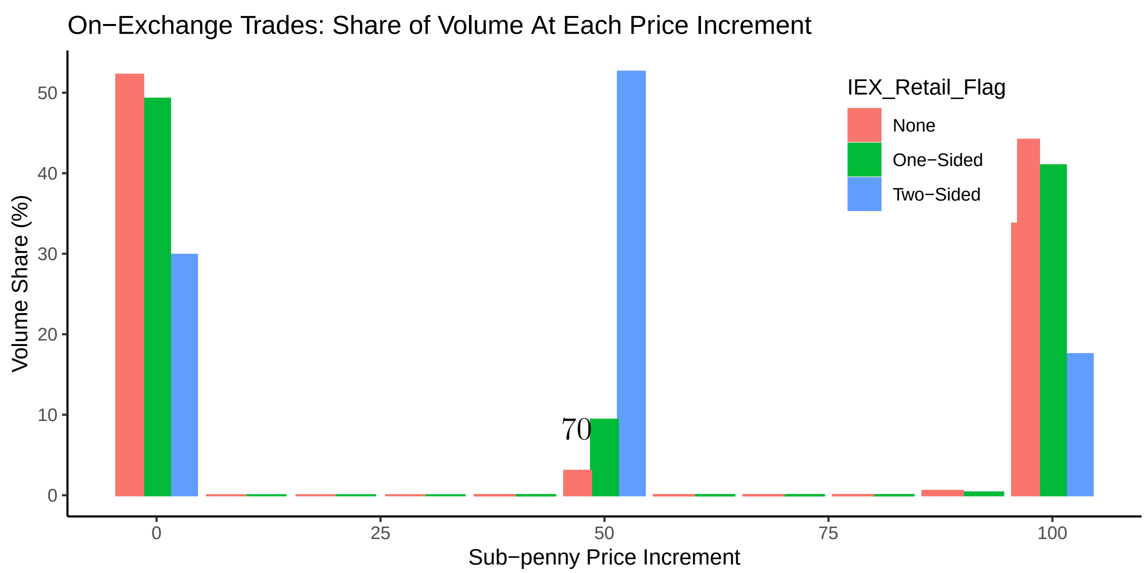


Figure 10. Volume Share of Off-Exchange Sub-Penny Prices. For each possible sub-penny price increment, we plot the volume occurring at this price increment, as a percentage share of all volume occurring on that venue. We use data from trades occurring on January 3, 2022, and separate trade volume into three categories: trades occurring when the IEX RLP has no interest, trades occurring when the IEX RLP has one-sided interest, and trades occurring when the IEX RLP has two-sided interest. Trades are restricted to be 100 shares or less, as the RLP only indicates availability for at least 100 shares. 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



Appendix I.A.2. A microfoundation of inventory cost structure ζ_i

In our baseline model, we assume that the marginal inventory cost for market maker i to execute a sell order is

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

In this section, we provide a microfoundation of this formulation, and illustrate the relations between stock liquidity and cost parameters c_0, c_1 and c_2.

We consider a one-period framework, with time $t = 0, 1$. There are N market makers labeled by $i = 1, 2..N$. For any market maker i, if its total net long position is z_i from time 0 to time 1, then it incurs a total inventory cost

$$\frac{1}{2}\gamma z_i^2$$

during this time period, and thus the marginal cost of executing the sell order is γz_i.[29] We assume that at the beginning of time 0, each market maker i's net long position is y_i. The sell order is then assigned to one of the N market makers according to a trading mechanism (broker's routing or order-by-order auctions). If market maker i obtains the sell order, it has to execute the order by internalizing it, routing it to other market makers (inter-dealer market), or sending it to the exchange. Right after market maker i receives the sell order, with probability $\alpha \in (0, 1)$, an inventory shock arrives, the market maker can not internalize the order, and has to either send the order to the exchange or execute it through the inter-dealer market. With probability η, there is active trading of the stock on the exchange, and market maker i can send the order to the exchange and close the position at cost $\bar{s}$. With probability $(1 - \eta)$, the market maker i can only send the order (randomly) to another market maker j. In this case, the cost is

$$\gamma_0 + \gamma y_j$$

where γ_0 is the fixed cost of connecting to another market maker and γy_j is the price charged

[29]This quadratic cost structure is commonly used in the literature (eg. Baldauf et al. (2022)).

by market maker j. For simplicity, we assume that market maker j offers competitive price γy_j, which is its marginal inventory cost. For simplicity, we make two implicit assumptions here. First, $\bar{s}$ is large enough, so it is always optimal for the market maker to internalize the order when the inventory shock is absent, and second, γ_0 is large enough so it is always optimal for the market maker to send the order to the exchange but not other market makers if possible.

Then the expected (marginal) cost of market maker i obtaining the sell order is

$$(1-\alpha)\gamma y_i + \alpha\left[\eta\bar{s} + (1-\eta)\frac{1}{N-1}\sum_{j\neq i}(\gamma_0 + \gamma y_j)\right].$$

The above cost can be rewritten as

$$[\alpha\eta\bar{s} + (1-\eta)\gamma_0] + \frac{1}{N}\left(\frac{\alpha(1-\eta)\gamma N}{N-1}\right)\sum_j y_j + \left((1-\alpha)\gamma - \frac{\alpha(1-\eta)\gamma}{N-1}\right)y_i.$$

Let

$$c_0 = \alpha\eta\bar{s} + (1-\eta)\gamma_0,$$

$$c_1 = \frac{\alpha(1-\eta)\gamma N}{N-1},$$

and

$$c_2 = (1-\alpha)\gamma - \frac{\alpha(1-\eta)\gamma}{N-1},$$

then the marginal cost for market maker i to execute the sell order is

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

Stock liquidity is linked to the parameter η in our microfoundation. When a stock is more liquid, it is more likely to have active trading on the exchange at that moment, and thus η will be larger. As a result, the ratio

$$\frac{c_2}{c_1} = \frac{N-1}{\alpha\gamma N}\left((1-\alpha)\frac{\gamma}{1-\eta} - \frac{\alpha\gamma}{N-1}\right)$$

will be larger. We utilize this interpretation in our discussions of model implications.

Appendix I.A.3. Omitted Proofs

Proof of Proposition 2

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

$$G_0\left(x|w\right) = \text{Prob}\left[\min_{-i} w_{-i} \le x | w_i = w\right]$$

and

$$g_0\left(x|w\right) = \frac{dG\left(x|w\right)}{dx}.$$

From this, it is straightforward that:

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

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

Let

$$
\begin{aligned}
v(x,w) =& \mathbb{E}\left[\zeta_i | \min_{-i} w_{-i} = x; w_i = w\right] \\
=& p_0 \mathbb{E}\left[\zeta_i | \min_{-i} w_{-i} = x; w_i = w; \exists j \neq i, w_j = y_j = x\right] \\
&+ (1-p_0)\, p_0 \mathbb{E}\left[c_i | \min_{-i} w_{-i} = x; w_i = y; \nexists j \neq i, w_j = y_j = x\right] \\
=& p_0 \left[c_0 + c_1 \left(\frac{x + (N-2)\, p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
&+ (1-p_0)\left[c_0 + c_1 \left(\frac{(N-1)\, p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
=& c_0 + c_1 \left(\frac{\left(p_0 x + (1-p_0)\, p_0 \frac{\frac{1}{2}+x}{2}\right) + (N-2)\, p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w.
\end{aligned}
$$

We focus on symmetric equilibria. Suppose all of market maker i's opponents use a continuous, increasing bid strategy

$$
B(w) = K_0 + K_1 w
$$

at time 0. When market maker i observes signal w and reports signal z, its expected profit is

$$
\begin{aligned}
U_i(z,w) &= \text{Prob}\left(z \leq \min_{-i} w_{-i} | w_i = w\right)\left[B(z) - \mathbb{E}\left(\zeta_i | z \leq \min_{-i} w_{-i}, w_i = w\right)\right] \\
&= [1 - G(z|w)]\left[B(z) - \frac{1}{1-G(z|w)} \int_z^1 g(x|w)\, v(x,w)\, dx\right] \\
&= [1 - G(z|w)]\, B(z) - \int_z^1 g(x|w)\, v(x,w)\, dx.
\end{aligned}
$$

Market maker i's marginal incentive is characterized by

$$
\begin{aligned}
\frac{\partial U_i(z,w)}{\partial z} &= -g(z|w)\,B(z) + (1 - G(z|w))\,B'(z) + g(z|w)\,v(z|w) \\
&= g(z|w)\left[-B(z) + \left(\frac{1 - G(z|w)}{g(z|w)}\right) B'(z) + v(z|w)\right] \\
&= g(z|w)\left[\begin{array}{c} -K_0 - K_1 z + \frac{\frac{1}{2}-z}{N-1}K_1 + c_0 \\ +c_1\left(\frac{\left(p_0 z + (1-p_0)p_0\frac{\frac{1}{2}+z}{2}\right) + (N-2)p_0\frac{\frac{1}{2}+z}{2} + p_0 w}{N}\right) + c_2 p_0 w \end{array}\right].
\end{aligned}
$$

We conjecture that in equilibrium we have

$$
\left.\frac{\partial U_i(z,w)}{\partial z}\right|_{z=w} = 0. \tag{I.A.3.1}
$$

This implies

$$
-(K_0 + K_1 w) + \frac{\frac{1}{2} - w}{N-1}K_1 + c_0 + c_1\left(\frac{\left(p_0 w + (1-p_0)\,p_0\frac{\frac{1}{2}+w}{2}\right) + (N-2)\,p_0\frac{\frac{1}{2}+w}{2} + p_0 w}{N}\right) + c_2 p_0 w = 0.
$$

Since the above condition holds for all w, then K_0,K_1 are solved by

$$
-K_0 + \frac{\frac{1}{2}K_1}{N-1} + c_0 + c_1\frac{N-2}{4N}p_0 + \frac{c_1(1-p_0)\,p_0}{4N} = 0,
$$

$$
-K_1 - \frac{K_1}{N-1} + c_2 p_0 + \frac{2c_1 p_0}{N} + \frac{c_1(N-2)\,p_0}{2N} + \frac{c_1(1-p_0)\,p_0}{2N} = 0.
$$

Then we get

$$
K_1 = \frac{N-1}{N}\left(c_2 p_0 + \frac{2c_1 p_0}{N} + \frac{c_1(N-2)\,p_0}{2N} + \frac{c_1(1-p_0)\,p_0}{2N}\right), \tag{I.A.3.2}
$$

$$
K_0 = c_0 + \frac{p_0}{4N^2}\left[\left(3 + N^2 - p_0 - Np_0\right)c_1 + 2Nc_2\right]. \tag{I.A.3.3}
$$

We also need to verify that condition (I.A.3.1) is a sufficient condition for optimization.

Note that $g\left(z|w\right) > 0$ and

$$-K_0 - K_1 z + \frac{\frac{1}{2} - z}{N-1} K_1 + c_0 + c_1 \left(\frac{\left(p_0 z + (1-p_0)\, p_0 \frac{\frac{1}{2}+z}{2}\right) + (N-2)\, p_0 \frac{\frac{1}{2}+z}{2} + p_0 w}{N} \right) + c_2 p_0 w$$

is linear in z, then it is clear that with (I.A.3.2) and (I.A.3.3), we must have that for all w,

$$\frac{\partial U_i\left(z,w\right)}{\partial z} < 0 \Longleftrightarrow z > w,$$

confirming that (I.A.3.1) is a sufficient condition for optimization.

Proof of Lemma 2

We introduce the random variable

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

and its CDF

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

and PDF

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

Then the total expected profit of market makers is

$$\begin{aligned}
W_M^{BR} =& \mathbb{E}\left\{\mathbb{E}\left[t\left(r\right)-c_0-c_1\frac{1}{N}\sum_{j=1}^{N}y_j-c_2y_i|w_i=\min_j w_j=r\right]\right\}\\
=&\mathbb{E}\left\{K_0+K_1r-c_0-c_1\left(\frac{p_0r+\left(N-1\right)p_0\left(\frac{1}{2}+r\right)\frac{1}{2}}{N}\right)-c_2p_0r\right\}\\
=&\mathbb{E}\left\{\frac{p_0}{4N^2}\left[2Nc_2\left(1-2r\right)+c_1\left(3-p_0\right)\left(1-2r\right)+c_1N\left(1-p_0\right)\left(1+2r\right)\right]\right\}\\
=&\int_{-\frac{1}{2}}^{\frac{1}{2}}\frac{p_0}{4N^2}\left[2Nc_2\left(1-2r\right)+c_1\left(3-p_0\right)\left(1-2r\right)+c_1N\left(1-p_0\right)\left(1+2r\right)\right]N\left(\frac{1}{2}-r\right)^{N-1}dr\\
=&\frac{p_0\left(2c_1-p_0c_1+Nc_2\right)}{N\left(1+N\right)}.
\end{aligned}$$

The expected total profit of investors W_I is

$$\begin{aligned}
W_I^{BR} &= -\mathbb{E}\left[K_0+K_1r|\min_i w_i=r\right]\\
&=-\int_{-\frac{1}{2}}^{\frac{1}{2}}\left(K_0+K_1r\right)N\left(\frac{1}{2}-r\right)^{N-1}dr\\
&=-\left[c_0+p_0\frac{2\left(2-p_0\right)c_1-\left(N-3\right)Nc_2}{2N\left(1+N\right)}\right],
\end{aligned}$$

and the total welfare W_{total} is

$$\begin{aligned}
W_{total}^{BR} &= \mathbb{E}\left\{\mathbb{E}\left[-c_0-c_1\frac{1}{N}\sum_{j=1}^{N}y_j-c_2y_i|w_i=\min_j w_j=r\right]\right\}\\
&=-\left(c_0-p_0\frac{N-1}{N+1}\frac{c_2}{2}\right).
\end{aligned}$$

Proof of Proposition 4

Since wholesalers are not able to observe the realization of $\tilde{c}_0$, they can condition their strategies only on the distributional information about $\tilde{c}_0$. We still focus on symmetric equilibria in this case, and conjecture that all wholesalers uses the same bidding strategy

$$\tilde{\beta}\left(y\right)=\tilde{k}_0+\tilde{k}_1y,$$

with $\tilde{k}_1 > 0$. Similar to our baseline model, the wholesaler with lowest signal realization obtains the order in equilibrium. We follow the proof of Proposition 2, notably, the function $\tilde{v}(x,w) = \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} w_{-i} = x; w_i = w\right]$ now becomes

$$\begin{aligned}
\tilde{v}(x,w) =& \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} w_{-i} = x; w_i = w\right] \\
=& p_0 \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} w_{-i} = x; w_i = w; \exists j \neq i, w_j = y_j = x\right] \\
&+ (1-p_0) p_0 \mathbb{E}\left[\tilde{\zeta}_i | \min_{-i} w_{-i} = x; w_i = y; \nexists j \neq i, w_j = y_j = x\right] \\
=& p_0 \left[c_0 + c_1 \left(\frac{x + (N-2) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
&+ (1-p_0) \left[c_0 + c_1 \left(\frac{(N-1) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
=& c_0 + c_1 \left(\frac{\left(p_0 x + (1-p_0) p_0 \frac{\frac{1}{2}+x}{2}\right) + (N-2) p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w,
\end{aligned}$$

which implies that

$$\tilde{v}(x,w) = v(x,w).$$

The rest of the proof follows the proof of Proposition 2. So the equilibrium strategy is the same as that in the baseline model.

Similarly, note that

$$\mathbb{E}(\tilde{c}_0) = c_0,$$

the proof of welfare computation follows our proof of Lemma 2, and thus all welfare outcomes are the same as that in our baseline model.

Proof of Proposition 8

Since the equilibrium of broker's routing is the same as that in the baseline model, we have

$$\tilde{W}^{BR}_{total} = -\left(c_0 - p_0 \frac{N-1}{N+1} \frac{c_2}{2}\right),$$

$$\tilde{W}_I^{BR} = -\left[c_0 + p_0 \frac{2(2-p_0)c_1 - (N-3)Nc_2}{2N(1+N)}\right],$$

and

$$\tilde{W}_W^{BR} = \frac{p_0(2c_1 - p_0c_1 + Nc_2)}{N(1+N)}.$$

For order-by-order auctions, the welfare outcomes are

$$\begin{aligned}\tilde{W}_{total}^{OBO} =& -\frac{1}{2}\left(c_0 + \delta_c - \frac{N+N_0-1}{N+N_0+1}\frac{c_2}{2}\right) - \frac{1}{2}\left(c_0 - \delta_c - \frac{N_0-1}{N_0+1}\frac{c_2}{2}\right)\\ =& - c_0 + \frac{1}{2}\left(\frac{N+N_0-1}{N+N_0+1} + \frac{N_0-1}{N_0+1}\right)\frac{c_2}{2},\end{aligned}$$

$$\begin{aligned}\tilde{W}_I^{OBO} =& -\frac{1}{2}\left[c_0 + \delta_c + \frac{1}{(N+N_0)(N+N_0+1)}c_1 - \frac{N+N_0-3}{2(N+N_0+1)}c_2\right]\\ & -\frac{1}{2}\left[c_0 - \delta_c + \frac{1}{N_0(N_0+1)}c_1 - \frac{N_0-3}{2(N_0+1)}c_2\right]\\ =& - c_0 - \frac{1}{2}\left[\frac{1}{(N+N_0)(N+N_0+1)}c_1 - \frac{N+N_0-3}{2(N+N_0+1)}c_2 + \frac{1}{N_0(N_0+1)}c_1 - \frac{N_0-3}{2(N_0+1)}c_2\right]\end{aligned}$$

and

$$\tilde{W}_W^{OBO} = \frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{c_1}{N+N_0} + c_2\right).$$

First,

$$\begin{aligned}& \tilde{W}_{total}^{OBO} > \tilde{W}_{total}^{BR}\\ \iff & - c_0 + \frac{1}{2}\left(\frac{N+N_0-1}{N+N_0+1} + \frac{N_0-1}{N_0+1}\right)\frac{c_2}{2} > -\left(c_0 - p_0\frac{N-1}{N+1}\frac{c_2}{2}\right)\\ \iff & \frac{1}{2}\left(\frac{N+N_0-1}{N+N_0+1} + \frac{N_0-1}{N_0+1}\right)\frac{c_2}{2} > p_0\frac{N-1}{N+1}\frac{c_2}{2}.\end{aligned}$$

Then LHS of the above condition is increasing in N_0, let $\underline{N}_0$ be the solution of

$$\frac{1}{2}\left(\frac{N+\underline{N}_0-1}{N+\underline{N}_0+1} + \frac{\underline{N}_0-1}{\underline{N}_0+1}\right)\frac{c_2}{2} = p_0\frac{N-1}{N+1}\frac{c_2}{2}, \tag{I.A.3.4}$$

then

$$N_0 > \underline{N}_0 \Longleftrightarrow \tilde{W}^{OBO}_{total} > \tilde{W}^{BR}_{total}.$$

Second,

$$\begin{aligned}
&\tilde{W}^{BR}_{W} < \tilde{W}^{OBO}_{W} \\
\Longleftrightarrow &\frac{p_0\left(2c_1 - p_0c_1 + Nc_2\right)}{N\left(1+N\right)} < \frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{c_1}{N+N_0} + c_2\right) \\
\Longleftrightarrow &\frac{p_0\left(2 - p_0 + N\frac{c_2}{c_1}\right)}{N\left(1+N\right)} < \frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{1}{N+N_0} + \frac{c_2}{c_1}\right) \\
\Longleftrightarrow &\frac{p_0\left(2-p_0\right)}{N\left(1+N\right)} + \frac{p_0}{1+N}\frac{c_2}{c_1} < \frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1}\left(\frac{1}{N+N_0} + \frac{c_2}{c_1}\right) \\
\Longleftrightarrow &\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}\right)\frac{c_2}{c_1} > -\left(\frac{1}{2}\frac{N}{\left(N+N_0\right)^2}\frac{1}{N+N_0+1} - \frac{p_0\left(2-p_0\right)}{N\left(1+N\right)}\right) \\
\Longleftrightarrow &\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}\right)\frac{c_2}{c_1} > -\frac{1}{N+N_0}\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{\left(1+N\right)}\frac{\left(2-p_0\right)\left(N+N_0\right)}{N}\right).
\end{aligned}$$

Since

$$\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N} > \frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{\left(1+N\right)}\frac{\left(2-p_0\right)\left(N+N_0\right)}{N},$$

we know that

$$\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}\right)\frac{c_2}{c_1} > -\frac{1}{N+N_0}\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{\left(1+N\right)}\frac{\left(2-p_0\right)\left(N+N_0\right)}{N}\right)$$

is equivalent to

$$\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N} > 0$$

and

$$\frac{c_2}{c_1} > \frac{-\frac{1}{N+N_0}\left(\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{(1+N)}\frac{(2-p_0)(N+N_0)}{N}\right)}{\frac{1}{2}\frac{N}{N+N_0}\frac{1}{N+N_0+1} - \frac{p_0}{1+N}}.$$

Finally,

$$
\begin{aligned}
&\tilde{W}_I^{BR} < \tilde{W}_I^{OBO} \\
\Longleftrightarrow & -\left[c_0 + p_0 \frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)}\right] \\
& < -c_0 - \frac{1}{2}\left[\frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)}c_1 - \frac{N+N_0-3}{2\left(N+N_0+1\right)}c_2 + \frac{1}{N_0\left(N_0+1\right)}c_1 - \frac{N_0-3}{2\left(N_0+1\right)}c_2\right] \\
\Longleftrightarrow & p_0 \frac{2\left(2-p_0\right)c_1 - \left(N-3\right)Nc_2}{2N\left(1+N\right)} \\
& > \frac{1}{2}\left[\frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)}c_1 - \frac{N+N_0-3}{2\left(N+N_0+1\right)}c_2 + \frac{1}{N_0\left(N_0+1\right)}c_1 - \frac{N_0-3}{2\left(N_0+1\right)}c_2\right] \\
\Longleftrightarrow & p_0 \frac{2\left(2-p_0\right) - \left(N-3\right)N\frac{c_2}{c_1}}{N\left(1+N\right)} \\
& > \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)} - \frac{N+N_0-3}{2\left(N+N_0+1\right)}\frac{c_2}{c_1} + \frac{1}{N_0\left(N_0+1\right)} - \frac{N_0-3}{2\left(N_0+1\right)}\frac{c_2}{c_1} \\
\Longleftrightarrow & p_0 \frac{2\left(2-p_0\right)}{N\left(1+N\right)} - p_0\left(1-\frac{4}{N+1}\right)\frac{c_2}{c_1} \\
& > \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)} - \frac{1}{2}\left(1-\frac{4}{N+N_0+1}\right)\frac{c_2}{c_1} + \frac{1}{N_0\left(N_0+1\right)} - \frac{1}{2}\left(1-\frac{4}{N_0+1}\right)\frac{c_2}{c_1} \\
\Longleftrightarrow & \left(1-p_0+\frac{4p_0}{N+1}-\frac{2}{N+N_0+1}-\frac{2}{N_0+1}\right)\frac{c_2}{c_1} > \frac{1}{\left(N+N_0\right)\left(N+N_0+1\right)} + \frac{1}{N_0\left(N_0+1\right)} - \frac{2p_0\left(2-p_0\right)}{N\left(1+N\right)}.
\end{aligned}
\tag{I.A.3.5}
$$

Proof of Proposition 9

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

$$G_0\left(x|w\right) = \text{Prob}\left[\min_{-i} w_{-i} \leq x | w_i = w\right]$$

and

$$g_0\left(x|w\right) = \frac{dG\left(x|w\right)}{dx}.$$

It is easy to show

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

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

Let

$$
\begin{aligned}
v(x,w) =& \mathbb{E}\left[\zeta_i | \min_{-i} w_{-i} = x; w_i = w\right] \\
=& p_0 \mathbb{E}\left[\zeta_i | \min_{-i} w_{-i} = x; w_i = w; \exists j \neq i, w_j = y_j = x\right] \\
& + (1-p_0)\, \mathbb{E}\left[c_i | \min_{-i} w_{-i} = x; w_i = y; \nexists j \neq i, w_j = y_j = x\right] \\
=& p_0 \left[\mathbb{E}(c_0) + \mathbb{E}(c_1)\left(\frac{x + (N-2)\, p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + \mathbb{E}(c_2)\, p_0 w\right] \\
& + (1-p_0)\left[c_0 + c_1 \left(\frac{(N-1)\, p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + c_2 p_0 w\right] \\
=& \bar{c}_0 + \bar{c}_1 \left(\frac{\left(p_0 x + (1-p_0)\, p_0 \frac{\frac{1}{2}+x}{2}\right) + (N-2)\, p_0 \frac{\frac{1}{2}+x}{2} + p_0 w}{N}\right) + \bar{c}_2 p_0 w.
\end{aligned}
$$

We focus on symmetric equilibria. Suppose all of market maker i's opponents use a continuous, increasing bid strategy

$$\bar{B}(w) = \bar{K}_0 + \bar{K}_1 w$$

at time 0. When market maker i observes signal w and reports signal z, its expected profit is

$$
\begin{aligned}
U_i(z,w) &= \text{Prob}\left(z \leq \min_{-i} w_{-i} | w_i = w\right)\left[\bar{B}(z) - \mathbb{E}\left(\zeta_i | z \leq \min_{-i} w_{-i}, w_i = w\right)\right] \\
&= [1 - G(z|w)]\left[\bar{B}(z) - \frac{1}{1 - G(z|w)} \int_z^1 g(x|w)\, v(x,w)\, dx\right] \\
&= [1 - G(z|w)]\, \bar{B}(z) - \int_z^1 g(x|w)\, v(x,w)\, dx.
\end{aligned}
$$

Market maker i's marginal incentive is characterized by

$$\begin{aligned}
&\frac{\partial U_i(z,w)}{\partial z} \\
=&-g(z|w)\bar{B}(z)+(1-G(z|w))\bar{B}'(z)+g(z|w)v(z|w) \\
=&g(z|w)\left[-\bar{B}(z)+\left(\frac{1-G(z|w)}{g(z|w)}\right)\bar{B}'(z)+v(z|w)\right] \\
=&g(z|w)\left[-\bar{K}_0-\bar{K}_1 z+\frac{\frac{1}{2}-z}{N-1}\bar{K}_1+c_0+c_1\left(\frac{\left(p_0 z+(1-p_0)p_0\frac{\frac{1}{2}+z}{2}\right)+(N-2)p_0\frac{\frac{1}{2}+z}{2}+p_0 w}{N}\right)+c_2 p_0 w\right].
\end{aligned}$$

We conjecture that in equilibrium we have

$$\left.\frac{\partial U_i(z,w)}{\partial z}\right|_{z=w}=0. \tag{I.A.3.6}$$

This implies

$$-\left(\bar{K}_0+\bar{K}_1 w\right)+\frac{\frac{1}{2}-w}{N-1}\bar{K}_1+\bar{c}_0+\bar{c}_1\left(\frac{\left(p_0 w+(1-p_0)p_0\frac{\frac{1}{2}+w}{2}\right)+(N-2)p_0\frac{\frac{1}{2}+w}{2}+p_0 w}{N}\right)+\bar{c}_2 p_0 w=0.$$

Since the above condition holds for all w, then $\bar{K}_0$,$\bar{K}_1$ are solved by

$$-\bar{K}_0+\frac{\frac{1}{2}\bar{K}_1}{N-1}+c_0+c_1\frac{N-2}{4N}p_0+\frac{c_1(1-p_0)p_0}{4N}=0$$

$$-\bar{K}_1-\frac{\bar{K}_1}{N-1}+c_2 p_0+\frac{2c_1 p_0}{N}+\frac{c_1(N-2)p_0}{2N}+\frac{c_1(1-p_0)p_0}{2N}=0$$

Then we get

$$\bar{K}_1=\frac{N-1}{N}\left(\bar{c}_2 p_0+\frac{2\bar{c}_1 p_0}{N}+\frac{\bar{c}_1(N-2)p_0}{2N}+\frac{\bar{c}_1(1-p_0)p_0}{2N}\right), \tag{I.A.3.7}$$

$$K_0=\bar{c}_0+\frac{p_0}{4N^2}\left[\left(3+N^2-p_0-Np_0\right)\bar{c}_1+2N\bar{c}_2\right]. \tag{I.A.3.8}$$

We also need to verify that condition (I.A.3.6) is a sufficient condition for optimization.

Note that $g\left(z|w\right) > 0$ and

$$-\bar{K}_0 - \bar{K}_1 z + \frac{\frac{1}{2} - z}{N-1}\bar{K}_1 + c_0 + c_1\left(\frac{\left(p_0 z + (1-p_0)\, p_0 \frac{\frac{1}{2}+z}{2}\right) + (N-2)\, p_0 \frac{\frac{1}{2}+z}{2} + p_0 w}{N}\right) + c_2 p_0 w$$

is linear in z, then it is clear that with (I.A.3.7) and (I.A.3.8), we must have that for all w,

$$\frac{\partial U_i\left(z, w\right)}{\partial z} < 0 \Longleftrightarrow z > w,$$

confirming that (I.A.3.6) is a sufficient condition for optimization.