Gone in Sixty Seconds: The Cost of Trading in Long Queues

Elaine Wah and Stan Feldman

IEX*

September 21, 2018

Abstract

The maker-taker pricing model, which pays market participants a rebate for providing liquidity, can lead to long queues at the exchanges employing this fee structure. But some participants may be able to get better queue position than others: high-speed traders can buy speed and data advantages in order to join the queue immediately, whereas slower investor orders are relegated to the back of the line. We analyze publicly available Daily TAQ data to estimate the costs of trading near or at the end of a long queue. By using aggregate quoted size at trade time as a proxy for queue priority, we calculate the impact and scale of performance differences associated with trading in long lines, which our results suggest may impose significant costs on investors.

1. Introduction

Of the 13 U.S. equities exchanges, seven currently pay market participants a per-share rebate for providing liquidity.[1] This pricing paradigm has been shown to be connected to longer lines to trade [Battalio et al., 2016, Wah et al., 2017]. But some traders may be able to get better queue position than others: faster market participants can exploit their speed advantages to join the queue at a new price level with near immediacy. By the time slower investors get in line, they have to wait for the orders ahead of them to execute first, and as a result their performance suffers. In this paper, we analyze Daily TAQ data to demonstrate that performance of orders trading near or at the end of a long line is substantially worse, which suggests that maker-taker rebates may impose significant costs on investors.

U.S. stock exchanges employ a number of different pricing paradigms based on which party is adding or removing liquidity. Every trade involves two participants: a "maker" who posts a buy or sell order (and in doing so gets in line at the back of the queue at the given price level), and a "taker" who trades against the order posted by the maker, either as a seller or a buyer. *Maker-taker*

 Corresponding author: Elaine Wah. Mailing address: IEX Group, Inc., 4 World Trade Center, 150 Greenwich Street, 44th Floor, New York, NY 10007. Email: `elaine.wah@iextrading.com`.

[1] Based on the top-tier access fees/rebates for adding liquidity in Tape A securities executed at or above $1.

exchanges pay a rebate for adding liquidity and charge an access fee for removing liquidity, whereas *inverted* exchanges charge a fee for adding and pay a rebate for removing. *Flat-fee* exchanges charge fees for both adding and removing liquidity, and do not pay rebates.

Access fees and rebates have come under increased scrutiny recently: In March 2018, the U.S. Securities and Exchange Commission proposed a Transaction Fee Pilot[2] to study the impact of fee structure on routing behavior and market quality, an initiative already endorsed by the U.S. Treasury [2017]. These initiatives are important because exchange pricing models have the potential to not only create conflicts of interest between brokers and their customers, but also spawn longer lines at the National Best Bid and Offer (NBBO), which represents the best prices at which one can buy or sell across all exchanges [Wah et al., 2017].

To illustrate how these long lines might arise: A market participant who posts an order on a maker-taker exchange receives a rebate for doing so (if and when the trade happens). However, a participant trading against a posted order on a maker-taker exchange is charged a fee for this trade, whereas she would be paid a rebate for doing the same thing on an inverted exchange. As such, makers (or liquidity adders) inherently have an incentive to post orders on maker-taker exchanges, and takers (or liquidity removers) who need to complete a trade have an incentive to go to inverted exchanges. This misalignment of incentives can contribute to long lines on the maker-taker exchanges, because the potential counterparty is incentivized to trade elsewhere.

Due to the price-time priority rules prevalent in today's equity markets, establishing a position in these queues early is inherently advantageous [Moallemi and Yuan, 2017]. To compete accordingly requires the ability to receive and respond to information as quickly as possible, which a market participant may attain by paying for direct data feeds, co-locating their servers within an exchange's data center, or investing in more sophisticated technology. Acquiring an informational edge then allows these faster participants (i.e., high-frequency traders and market makers) to respond near instantaneously to market events such as quote changes. Lacking the speed advantages to respond immediately to NBBO changes, slower market participants (i.e., investors), have to wait at the back of long lines.

But what is the potential cost ultimately borne by investors of these long lines? To answer this question, we analyze Daily TAQ data to estimate the cost of trading near the end of a long line. We describe our dataset in Section 2 and our model and methodology in Section 3. We discuss our results in Section 4, and we conclude in Section 5.

2. Data

For our analysis, we use publicly available Daily TAQ data, which comprises trade and quote data with microsecond-level timestamps. Our dataset consists of trades and quotes from May 2018. We include data from 12 U.S. equities exchanges in our analysis. The maker-taker venues included are Arca, Cboe BZX, Cboe EDGX, Nasdaq, New York Stock Exchange (NYSE), and Nasdaq PHLX (PSX); the inverted or flat-fee exchanges are Nasdaq BX, Cboe BYX, Cboe EDGA, Investors Exchange (IEX), NYSE American (MKT), and NYSE National (NSX). As with Wah et al. [2017], we exclude the Chicago Stock Exchange (CHX) due to sample size and data robustness concerns.

We include only trades at the NBBO, which reflect the executions of the market participants in line at the best prices across all venues. As for the TAQ quote data, we apply the same filters as Wah et al. [2017], which include excluding locked and crossed markets, as well excluding abnormal

[2]See `https://www.sec.gov/rules/proposed/2018/34-82873.pdf`.

quotes (i.e., where the NBO is outside the range $\left[\frac{1}{3}NBB, 3\,NBB\right]$, where NBB is the National Best Bid and NBO is the National Best Offer).

3. Methodology

Since TAQ data comprises only publicly reported trade and quote data, and does not include any information about the original underlying orders, it is impossible to ascertain actual order priority for an observed execution from TAQ. However, we can use aggregate quoted size for a given symbol at trade time as a proxy for queue priority:

- Trades when the aggregate quoted size is large (relative to the symbol's typical queue size) are likely executing at the front of a long line.
- Trades when the aggregate quoted size is small (relative to the symbol's typical queue size) are likely executing at the end of the line, after everything ahead in the line has already been exhausted.

To estimate the costs to investors of trading in long lines, we group trades based on the aggregate quoted size at the NBBO during the time of trade. More specifically, we group executions of resting buy (sell) orders at the NBB (NBO) based on the total size available across all exchanges at the NBB (NBO).

We discuss how we assign trades to weighted deciles based on aggregate quoted size in Section 3.1. To compare the performance of trades within each group across different types of exchanges, we compute trade markouts as described in Section 3.2.

3.1. Weighted deciles

Since traded volume can vary significantly by execution, we group trades into weighted deciles by volume in order to ensure that we can compare performance on a volume-equivalent basis. Standard deciles ensure only the same total number of observations per group; in contrast, weighted deciles ensure the same total weight—in our case, volume—per group.

To avoid implicit bias from preexisting orderings such as time of day, we analyze a random permutation of the trades dataset, analogous to shuffling a deck of cards. Note that this does not alter the decile to which a trade belongs, except in the instances where multiple trades near and around the volume threshold between two deciles have the same total quoted size.

Given N trades in symbol s on a given exchange on a given date, we first sort the trades in descending order by aggregate quoted size on the side in question, that is, from high to low based on the total size available at the NBB (NBO) for a trade at the NBB (NBO). If the sorted trades have trade sizes $q_1, q_2, q_3, \ldots, q_N$, let Q_s represent the total executed volume for symbol s, where $Q_s = \sum_{i=1}^{N} q_i$.

There are ten deciles, so each decile should comprise $\frac{1}{10}Q_s$ shares. We assign the kth trade to the first decile as long as the cumulative volume up to the kth trade is less than or equal to the volume per decile, or $\frac{1}{10}Q_s$. That is, every trade up to and including the kth trade is assigned to the first decile if the following holds:

$$\sum_{i=1}^{k} q_i \leq \frac{1}{10}Q_s \tag{1}$$

More generally, we assign the kth trade to the nth decile if the following condition holds:

$$\frac{n-1}{10}Q_s < \sum_{i=1}^{k} q_i \leq \frac{n}{10}Q_s \tag{2}$$

Our method does not split up trades that straddle the threshold between deciles, so the total volume per decile is not necessarily exactly the same.

3.2. Markouts

We measure performance via trade markouts, or realized spread, which are a standard in both industry and the academic literature. Markouts compare the price at execution to the midpoint of the market at some specified future time after the trade. More positive markouts reflect greater potential profit after the trade, whereas negative markouts reflect buyer's or seller's remorse—i.e., when the price has gone down (up) after the trade for a buyer (seller). These adverse selection costs arise when informed traders sell to (buy from) a resting buy (sell) order right before prices fall (rise).

Trade markouts are typically restricted to executions at the NBB or NBO because it is not always possible to determine the direction of an execution (i.e., whether the liquidity remover was a buyer or seller) happening inside the NBBO. This constraint does not affect our model, however, as we are only concerned with executions at the NBB or NBO, as these reflect trades of the participants waiting in line at the inside quote.

Trade markouts are measured from the perspective of the resting order, or the participant waiting in the line to trade. Given the NBBO midpoint $M_t = \frac{1}{2}(NBB_t + NBO_t)$ for a given symbol at time t, we define the markout for a trade i that executed at price $p_{i,t}$ at time t as follows:

$$\delta_{\text{markout}} = \begin{cases} M_{t+\tau} - p_{i,t} & \text{for buy orders} \\ p_{i,t} - M_{i,t+\tau} & \text{for sell orders} \end{cases} \tag{3}$$

where $\tau > 0$ is some fixed time interval. We volume-weight markouts by the shares executed.

We also compute volume-weighted relative markouts. A relative markout is the trade markout defined above divided by the NBBO midpoint at the time of trade. More formally, we define the relative markout $\delta_{\%\text{markout}}$ for a trade i that executed at time t as follows:

$$\delta_{\%\text{markout}} = \begin{cases} \dfrac{M_{t+\tau} - p_{i,t}}{M_t} & \text{for buy orders} \\ \dfrac{p_{i,t} - M_{t+\tau}}{M_t} & \text{for sell orders} \end{cases} \tag{4}$$

4. Results

Overall, our results demonstrate the drastic difference in performance between trading at the front of the line versus the back. This is consistent across all exchanges, as evidenced in Figure 1. Notably, the most negative markouts are on the largest maker-taker exchanges such as NYSE, Arca, and Nasdaq, reflecting the greatest degree of adverse selection; in contrast, the inverted/flat-fee venues have universally positive markouts for the first 5 deciles. The plot only shows trade markouts at the 1-minute mark, but our qualitative results are robust given other settings of τ (including 1ms, 10ms, 100ms, 1s, 5s, 10s, 20s, 30s, and 5min).

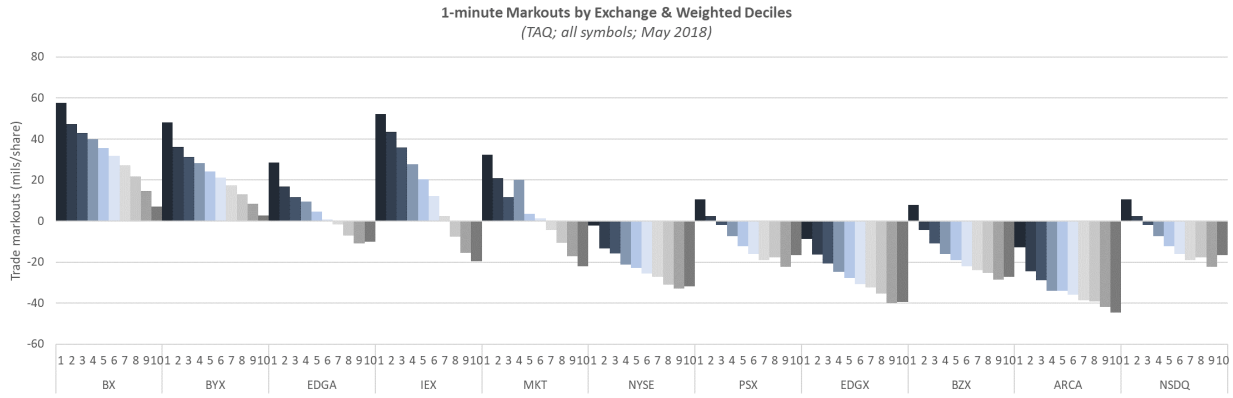

Fig. 1. Markouts by exchange and decile, over all symbols. Deciles are labeled from 1 through 10 for each exchange. The first decile for each exchange is in dark blue; the tenth decile is in gray. NYSE National is excluded here due to low volume during the period in question.

When we group the results by exchange type, as in Figure 2, a clear pattern emerges. Trading on inverted/flat-fee exchanges is associated with better performance, whereas trading on the maker-taker exchanges is worse regardless of place in the queue. The slight uptick in the 10th decile relative to the 9th decile on maker-taker exchanges may be because relatively low aggregated quoted size at the inside can potentially also reflect instances before the queue has fully formed, when only the fastest market participants are in line. Trades during such instances are likely rare, however, given that the potential counterparties would need to be equally fast (in order to have seen the queue start to form) and willing to cross the spread to execute.

We observe that performance in the last decile of an inverted/flat-fee venue is better than performance in the first decile of a maker-taker venue. This suggests that exchanges are generally accessed in order of the cost to remove liquidity, since the inverted/flat-fee venues pay a rebate (or charge a low fee) to liquidity removers whereas the maker-taker venues typically charge the maximum fee allowed (30 mils per share). In other words, a non-trivial portion of order routing is done on a cost-effective basis. Furthermore, our results indicate that markouts associated with the last 5 deciles on maker-taker venues are higher in magnitude than the highest-tier rebate paid by these exchanges. This suggests that the sub-optimal outcomes associated with being among the last to trade on these venues are not necessarily counterbalanced by the rebate payment.

So what do these results mean in terms of the costs imposed on investors for trading near or at the end of long queues? To estimate the cost to investors, we multiply the by-decile performance on maker-taker exchanges by the volume traded on those venues. We assess performance on a decile-by-decile basis to capture the relative differences in speed and access across brokers, and to err on the conservative side. An investor order placed at the end of a long line on a maker-taker exchange is potentially costing the investor a better execution elsewhere, but orders at the back of the line (represented by trades in the 10th decile) on maker-taker exchanges are also less likely to be able to compete for optimal queue position on other venues.

Our results by decile are in Table 1, which shows the annualized cost estimates on a per-decile basis. We expect the market participant composition of each decile to vary significantly: the lower deciles (the front of the queue) are more likely to include high-speed traders, whereas the higher deciles (the back of the queue) are more likely to include long-term investors. However, we cannot determine with certainty how much of each decile is comprised of investor orders. Another

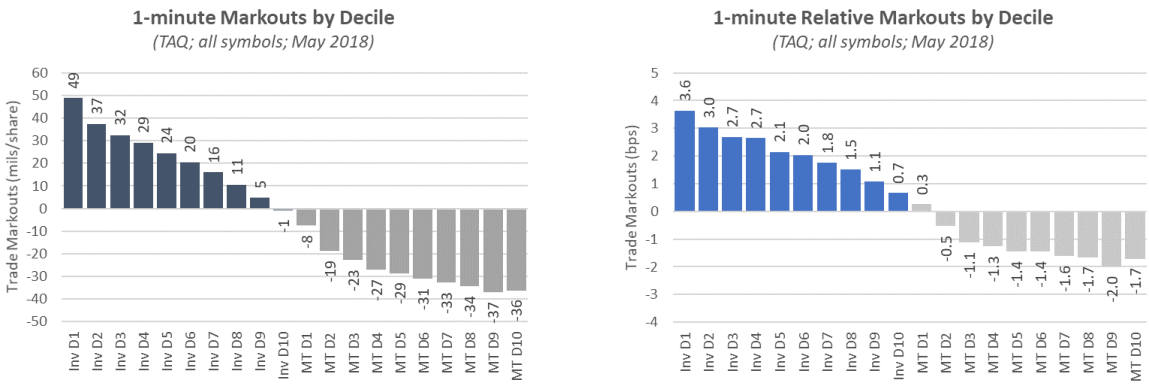

Fig. 2. Markouts by decile and by exchange type, over all symbols. Trade markouts (mils per share) are on the left, with relative trade markouts (bps) on the right. Deciles are labeled from D1 (1st decile) through D10 (10th decile). Inverted/flat-fee ("Inv") results are shown in blue and maker-taker ("MT") results are in gray.

consideration in evaluating these results is whether the rebate is passed back to a firm adding liquidity on maker-taker exchanges, as the rebate payment may partially or fully compensate any performance degradation.[3] For these reasons, we do not provide a singular estimate of "total harm." Nonetheless, our results unequivocally show that performance on maker-taker venues is substantially worse at the back of the queue—in just the first 60 seconds after trading! As such, any investor orders sent to maker-taker exchanges are subject to significant potential cost, especially if the rebate is not passed through to the end investor.

We note that our results presented here are generally in line with prior estimates.[4] A KCG analysis found a 4.5 basis points difference in performance between orders at the very front versus the back of the queue on Nasdaq and Nasdaq BX [Mackintosh, 2014]. Using our methodology, the difference in 30-second relative markouts between the 1st decile for inverted/flat-fee exchanges and the 10th decile for maker-taker exchanges is 5.3 bps, as seen in Figure 2.

[3]We note that quantitative hedge funds typically employ "cost-plus" pricing models in which the rebate is passed through to the end client; however, we lack the data to estimate how much volume in each decile is due to these funds.

[4]We also validated our results by analyzing a day's worth of order-by-order data: we assigned trades into weighted deciles by each trade's original order priority, and the qualitative results were consistent with those presented here. However, due to data distribution restrictions, we are unable to report the order-by-order results here.

Table 1: Estimated costs of trading by decile. We annualize based on the 1-minute markouts (mils per share) on maker-taker exchanges multiplied by the average daily volume (shares) on maker-taker exchanges in the given decile.

Decile	Maker-Taker	M-T Volume	Daily Cost ($)	Annualized Cost ($)
1	(7.6)	187,512,946	$142,497	$35,909,346
2	(18.8)	196,829,051	$370,030	$93,247,502
3	(22.8)	197,260,955	$449,410	$113,251,416
4	(26.9)	197,450,345	$531,411	$133,915,506
5	(28.5)	197,445,543	$563,510	$142,004,515
6	(30.9)	197,570,169	$610,030	$153,727,598
7	(32.8)	197,432,795	$647,161	$163,084,550
8	(34.4)	197,640,815	$680,049	$171,372,351
9	(36.9)	197,966,748	$731,065	$184,228,280
10	(36.1)	205,614,572	$743,158	$187,275,740

5. Conclusion

In this study, we analyzed Daily TAQ data to estimate the impact to investors of trading in the back half of the queue at the National Best Bid and Offer. We use aggregate quoted size as a proxy for queue priority: trades executing when the quoted size is large are likely occurring at the front of a long line, whereas trades executing when the quoted size is small are likely occurring near the back of the line, when the rest of the displayed quote has already been exhausted. Lacking the speed and data advantages purchased by faster market participants, investors are unable to join the queue immediately. By the time investors get in line, the queue is already long, which could ultimately result in substantial and unnecessary losses.

Our model only captures the costs of trading near or at the end of the line, but investors also suffer the opportunity cost of either canceling or simply *not* trading, due to the length of the line. By the time slower investors get in line, they have to wait for the orders ahead of them to execute first—which reduces their likelihood of trading. As such, our estimate likely understates the total cost of trading in long queues. Since TAQ data only includes trades and quotes, an estimate of opportunity cost to investors necessitates more granular data.

Maker-taker and inverted pricing models have made it more difficult for buyers and sellers to meet, because they create economic incentives for different—but mutually dependent—types of activity to be allocated to different exchanges, which can lead to unnecessary intermediation. Oftentimes, makers will only meet takers when the urgency to trade is great enough to incentivize forgoing the rebate (or once orders on venues paying rebates to liquidity removers have been exhausted). For instance, a market participant may want to trade against all displayed shares available, and will thus execute against posted orders at any venue with available liquidity, despite the fee for taking on maker-taker venues.

The exchange pricing models entrenched in today's U.S. equity market structure have ultimately created a system in which investor orders are being placed at the end of long lines, and trading is fragmented across multiple venues. And the significant performance disparity between trading at the front of the line versus the back has placed a premium on high-speed market data and connectivity. To remain competitive, market participants have no choice but to purchase data and connectivity from multiple exchanges—thereby perpetuating an ecosystem in which exchanges profit at the expense of investors.

Further work will be necessary to determine the full impact of access fees and rebates on investors, but the SEC's proposed Transaction Fee Pilot, most notably with its zero-rebate bucket, is a much-needed step towards eliminating the conflicts of interest present today and further safeguarding investors.

References

Robert Battalio, Shane A Corwin, and Robert Jennings. Can brokers have it all? On the relation between make-take fees and limit order execution quality. *Journal of Finance*, 71(5):2193–2238, 2016.

Phil Mackintosh. The need for speed: It's important, even for VWAP strategies. Technical report, KCG, 2014.

Ciamac C. Moallemi and Kai Yuan. A model for queue position valuation in a limit order book. Technical report, Columbia Business School Research Paper No. 17-70, 2017.

U.S. Treasury. A financial system that creates economic opportunities: Capital markets. Available at: `https://www.treasury.gov/press-center/press-releases/Documents/A-Financial-System-Capital-Markets-FINAL-FINAL.pdf`, 2017.

Elaine Wah, Stan Feldman, Francis Chung, Allison Bishop, and Daniel Aisen. A comparison of execution quality across U.S. stock exchanges. *SSRN Electronic Journal*, pages 1–55, 2017.