The Value of Off-Exchange Data*

Thomas Ernst,† Jonathan Sokobin^ and Chester Spatt§

December 28, 2021

Abstract

Exploiting the structure of geographic latencies, we study the effect of trade reporting of *off-exchange* equity transactions and contrast that with reporting of *exchange* trading. Publication of *off-exchange* transactions by the Securities Information Processor (SIP), leads to a sharp burst in trading and quoting activity, suggesting that market participants learn from those reports, with their unique information content lingering throughout the lengthy reporting process. In contrast, there is no spike in response to SIP publication of exchange trading, but instead an earlier spike that reflects the response to the near-immediate reporting from proprietary feeds. Due to the varied locations of the off-exchange trade reporting facilities (TRFs), SIPs and exchanges, we use distinct geographical latencies to pinpoint the patterns. We document that realized spreads for the TRF-response trades are negative, consistent with these orders being informationally-motivated and contributing to price discovery.

Keywords: Market data; trades and quotes; high-frequency trading; off-exchange trading; Regulation NMS; proprietary data; post-trade opacity; microseconds

The views and opinions expressed are those of the authors and do not represent official policy of the Financial Industry Regulatory Authority (FINRA).

∗First Draft: January 2021. For helpful comments, we are thankful to Bobby Bartlett, Pete Kyle, Haoxiang Zhu, and seminar participants at the NYU Stern Microstructure Conference, the University of Maryland, and Carnegie Mellon University.
†University of Maryland, Robert H. Smith School of Business, ternst@umd.edu
^Financial Industry Regulatory Authority, Jonathan.Sokobin@finra.org
§Carnegie Mellon University, Tepper School of Business, cspatt@andrew.cmu.edu

I. Introduction

Two important and related trends in the equity markets raise questions about the information content and value of off-exchange trade information. First, is the rise of off-exchange trading--initially with ATSs' and subsequently with retail order flow being sent to internalizers.[1] Almost half of all U.S. equity trading volume now occurs off-exchange, with the off-exchange share increasing considerably during 2020. Second, data revenues are expanding rapidly, becoming a key component of an exchange's business model over the last five years.[2] The design of markets, embedded latencies, and the incentives created from monetizing the data flow through proprietary feeds has become a contentious and litigious issue, receiving considerable attention among market participants.[3] Though Reg NMS has been in force for about 15 years, the trading landscape continues to evolve with the joint growth of off-exchange trading volume and the growth of on-exchange data value.

This paper is one of the first to analyze the value of *off-exchange* data: what the value is, where it comes from, and how the value of off-exchange data differs from that of on-exchange data. Our work highlights the critical importance of post-trade transparency: while on-exchange trades are reported in tens of microseconds, off-exchange trades are reported thousands of microseconds after they occur. When off-exchange trades finally are reported, we document a sharp, rapid increase in trades and quotes in response. We identify these trade reactions by exploiting geographic variation

[1] In 2015, the percentage of NMS equities traded off-exchange (either through an ATS or internalizing firm) represented 35.4% of trading volume and by 2019, it had risen to 37.2% See FINRA's Industry Snapshot, available at: https://www.finra.org/rules-guidance/guidance/reports-studies/2020-industry-snapshot/market-data. More recently, it has been reported that the proportion of NMS equities trading off-exchange has exceeded 50%. See, e.g., "Rise of Retail Army Shrouds Half of U.S. Stock Trading" available at: https://www.bloomberg.com/news/articles/2021-02-17/rise-of-retail-army-shrouds-half-of-stock-trading-in-secrecy.

[2] Intercontinental Exchange, the owner of the New York Stock Exchange, reports data revenues of over $ 2.2 billion in 2019, compared to less than $ 1 billion in 2014. Nasdaq, Inc. reports data revenues of $ 779 million in 2019, up from $ 473 million in 2014.

[3] For example, see "Petition for Rulemaking Concerning Market Data Fees," December 6, 2017. https://www.sec.gov/rules/petitions/2017/petn4-716.pdf and also U.S. Securities and Exchange Commission, Ruling in Application of SIFMA for Review of Action Taken by NYSE Arca and Nasdaq, October 16, 2018, https://www.sec.gov/litigation/opinions/2018/34-84432.pdf, which was reversed by the District of Columbia Circuit Court of Appeals (June 5, 2020) at https://www.cadc.uscourts.gov/internet/opinions.nsf/127CE4C0762C082F8525857E00506366/$file/18-1292-1845826.pdf as well as "SEC Roundtable on Market Data Products, Market Access Services and Their Associated Fees," October 25 and 26, 2018, https://www.sec.gov/spotlight/equity-market-structure-roundtables/roundtable-market-data-market-access-102518-transcript.pdf and https://www.sec.gov/spotlight/equity-market-structure-roundtables/roundtable-market-data-market-access-102618-transcript.pdf.

in markets: across multiple distinct market facilities, the spike in trades and quotes lines up exactly with known geographic latencies. These races to respond to an off-exchange trade appear profitable: responses are overwhelmingly the same sign as the off-exchange trade they respond to, and the responses earn negative realized spreads on average. Following Boehmer, Jones, Zhang, and Zhang (2021) we divide dark trades into full penny price (institutional) and sub-penny price (retail) trades. Institutional trades are more numerous than retail trades, but the patterns in responses to the two groups are similar. Total response volumes are sizeable: we document $775 billion per year in quote reactions, and $65 billion per year in trade reactions for the stocks of our sample. While the latencies involved are shorter than the blink of an eye, the value of an advantage from first access to data can be considerable.

Off-exchange trades are reported in a two-stage process: the trade is first reported to a Trade Reporting Facility (TRF), and the trade is subsequently sent by the TRF to the Securities Information Processor (SIP), which broadcasts the trade to market participants more broadly. This differs from exchange executed trades, where at execution the trade is simultaneously reported to proprietary data feeds and the SIP. We empirically document sharp, rapid increases in trading and quoting activity in response to the SIP publication of off-exchange trades. Realized spreads for these trades are negative, consistent with the SIP broadcasts initiating a race between messages seeking to cancel existing quotes, and messages seeking to trade with existing quotes before they can be updated to reflect this new information and contribute to price discovery (e.g., the response orders are able to "pick off" stale prices, as in Foucault, Roell and Sandas (2003)). At each stage of the process, we examine market forces which are ultimately reflected in the cost to access market data and the payments made to traders through rebates and price improvement. We contrast the value of the trade report information inferred from the subsequent trading and quoting activity to the data fees because the data fees should reflect the value of the information to other traders. An interesting context that examines the value of early release of fundamental market information and how that can promote price discovery is Hu, Pan and Wang (2017).

We exploit several distinct sources of geographic latency, which allows us to pinpoint the market reaction to specific pieces of information. Within New Jersey, there are two TRFs, two SIPs, and three exchanges, giving twelve distinct pathways for information to reach the market. The patterns in market activity arise on all three of the major equity exchanges, and tightly align with known SIP-to-exchange geographic latencies. Along each of these twelve pathways, we see

a rapid increase in trade and quote activity at a specific exchange which precisely aligns with the time that the SIP publication of an off-exchange trade would reach that exchange. This set of twelve geographic pathways allows us to isolate the response to off-exchange trades. For example, if a trader places a simultaneous on-exchange and off-exchange trade, the information about the off-exchange trade will have a distinct path through the TRF and SIP to market participants, allowing us to separate the response to off-exchange information from any confounding with the release of the on-exchange information.

The difference in post-trade transparency between on-exchange and off-exchange trades is considerable. Academic focus on dark trading has highlighted the lack of pre-trade transparency.[4] Just as striking, however, is the delay in *post*-trade transparency, which arises from the trade-reporting latencies. For on-exchange trades, proprietary data feeds announce a trade publicly in less than 50 microseconds.[5] The message is also public: parties to the trade are notified of a successful trade at the same time that all market participants are informed. In contrast, the median dark trade is not reported to other market participants until 2,500 microseconds after the trade occurs.[6] Unlike the lit market, the parties to the off-exchange trade can be notified well before the rest of the market, and thus, have the potential opportunity to place additional orders before their trade is publicly revealed to other market participants, even those subscribing to the lowest-latency data.[7] Whether the TRF trades are included in the fastest feeds is a decision of the exchanges and endogenous.

Using trade timestamps from the exchange, the TRF, and the SIP allow us to explore the information flow at each step. Market activity occurs in response to information. For the

[4] Comerton-Forde and Putnins (2015) and Zhu (2014).

[5] Kim and Trepanier (2020) measure round-trip times on NASDAQ below 50 microseconds, implying a one-way journey of around 25 microseconds. CBOE reports a order-to-quote round-trip 78 microseconds on average, and 80% of all orders occurring with 73 microseconds: https://cdn.cboe.com/resources/features/Cboe_US-Equities_Latency-FactSheet.pdf
NYSE reports one-way latencies with medians all below 50 microseconds, available at: https://www.nyse.com/pillar

[6] In an earlier era there was asymmetric knowledge of market transactions that were handled by block trading desks, allowing the principal to obtain better execution by committing not to undertake additional trades and undercut the block trading firm (see Seppi (1990)), while the block desk is still working the acquired inventory position, referred to as "no bagging the Street." To a degree, this is a counterpart to off-exchange trading at present. Knowledge of the recent block execution provided an important advantage to the parties that had direct knowledge of that transaction (which was the underpinning of the "no bagging the Street" convention).

[7] We note that there is considerable variation among off-exchange trading venues. Depending on the platform of the trading venue or the broker, there may be variation in when a broker, or its ultimate customer, receives a trade confirmation.

exchanges, trade and quote activity occurs very shortly after the exchange trade, consistent with market participants subscribing to, and learning from, proprietary exchange feeds. When exchange trades are published by the SIP, there is little to no reaction to the information; presumably, market participants had already seen the information through proprietary feeds. In contrast, when an off-exchange trade arrives at a TRF, there is no resulting change in quoting activity. When that same off-exchange trade is published by the SIP, there is a sharp increase in trading and quoting activity. This response to the SIP publication of trades is specific to SIP reports of off-exchange trades. Market participants appear to learn about off-exchange trades from the SIP broadcasts, and place trades or quotes in response to this information.[8]

While the market is informed of off-exchange trades with a considerable delay, the participants in the off-exchange trade themselves will have knowledge of the trade before the broader market. The parties to an off-exchange trade are potentially aware of the trade thousands of microseconds before the trade is published. We show that there is a considerable elevation of exchange trading around the time these off-exchange trades take place, indicating that at least one party to an off-exchange trade (whether a buy-side investor or market maker) also trades on-exchange. We collect evidence for direct observation of two possible trading strategies: simultaneous or sequential market access. Under simultaneous access, a market participant trades in both dark and lit markets at the same time. Under sequential access, a market participant trades off-exchange, and then goes to exchanges to trade before the off-exchange trade is published. With this sequential approach to trading, a market participant can exploit the latency in trade reporting to access quotes on the exchange before those quotes can be revised in response to the new trade information. This is not possible for exchange trades, as those trades are reported to all participants at the same time.

We note that for exchange trading the pattern is very different—there the market reaction (spike in quoting and trading activity) is to the proprietary data feeds rather than to SIP publication, which is what we would have anticipated a priori. This highlights the potential value of such proprietary data feeds. By considering off-exchange trading, we introduce a somewhat different

[8] An additional (non-geographic) source of identification for our results is that the NYSE only started to include its off-exchange trades in some of its proprietary data feeds on April 29, 2019. We confirm that the response to the SIP publication seems largely unchanged and there is not a noticeable change in response to the trade report at the inclusion in the proprietary data.

regime and perhaps surprisingly, obtain quite different results about the role and value of proprietary vs. SIP data. It bears emphasis that the regulatory regimes in the exchange and off-exchange cases are different. Trade reporting is the responsibility of an exchange at which trading occurs, while a broker-dealer involved in an off-exchange trade has the trade reporting obligation and can direct the trade reporting to a particular TRF (so trading off-exchange then involves a greater time differential between trade reporting and trade execution). Bartlett and McCrary (2019) highlight that trading using exchange proprietary feeds against traders employing only SIP data feeds generates small profits. Consistent with this, we show that there is little trading or quoting response to the SIP publication of an exchange trade, but we also highlight the surprising fact that all traders learn about off-exchange trades through their SIP publication.

Strikingly, these reporting facilities for "off-exchange" trading are operated by the leading exchanges (which in principle could be a source of expertise and experience or a conflict of interest)[9] and as highlighted in Spatt (2021), on the exchange side there is the potential for the major exchanges to subsidize trading through very small or even negative net fees in order to enhance the value of the exchange market data (and connectivity) and overall profitability. Recent regulatory actions by the SEC have highlighted the potential for changes in the model for data provision with recent changes in the governance of the National Market Systems data plans (i.e., the current SIP) and a move towards facilitating competing consolidators.[10]

This paper offers a number of contributions. The paper highlights and focuses attention upon the importance of off-exchange trading and the value of its data. It casts in a broader context the role of different data and participants, helping to understand the relationship between the SIP and proprietary data, the possibility of a conflict associated with exchanges operating the Trade Reporting Facilities for non-exchange trading, and post-trade opaqueness of dark pools and off-

[9] Although likely not a key issue at the time that the TRFs were established in 2006, it is interesting that the Trade Reporting Facilities are operated by the major trading exchanges and key participants in the governance of the SIP (NYSE and NASDAQ) rather than FINRA or some other third party such as a technology firms or media companies. Arguably, the exchanges would have the greatest expertise and experience in trade reporting due to the obligations it had been fulfilling with respect to exchange trade reporting. As the relative importance of data revenues have increased over the time, the potential for a conflict of interest has also increased because of the potential for cross-price market effects between exchange and off-exchange reporting.

[10] The changes that the SEC adopted (see https://www.sec.gov/news/press-release/2020-311 and https://www.sec.gov/rules/final/2020/34-90610.pdf) have been challenged by the exchanges through a lawsuit to the District of Columbia Circuit Court of Appeals.

exchange trading more broadly.[11] On the latter front, our study highlights the faster effective response and reporting of exchange trading (in that the spike in trade responses occurs before the SIP) than for off-exchange trading, suggesting a potentially important proprietary advantage of off-exchange (dark) trading and the importance of reactions to dark trades. We also show the cross-venue activity of investors, with investors who trade off-exchange also trading on-exchange.

Much of the attention to opacity in off-exchange and dark pool trading has focused upon pre-trade opacity rather than the important opacity about relative trade reporting that emerges in the aftermath of trading through latency (despite post-trade reporting requirements in equity). The reporting of exchange trades actually is more rapid (by several milliseconds) than off-exchange trades. This ties closely to the impact of latency frictions upon price discovery in the market and subtle aspects of staleness in prices related to the distinction between exchange and off-exchange reporting and the geographic structure of reporting [Budish, Cramton and Shim (2015), Aquilina, Budish and O'Neill (2020), and Hu, Pan and Wang (2017)]. Hasbrouck (2019) highlights how price discovery measures change with the resolution of timestamps, and the importance of the proprietary data feeds for price discovery. Our work highlights the fact that while exchange trades are quickly reported, off-exchange trades, and the associated market reaction, are significantly delayed. The on-exchange portion of cross-venue activity of investors is reported earlier, which blunts the reaction to the off-exchange component as market participants have already seen part of the overall order.

Our analysis also has important ramifications for both the meaning of Best Execution and the measurement of the National Best Bid and Offer (NBBO) in that the data and information available to broker-dealers at various locations and the latency in the system are important to their execution obligations and strategies (the vantage point of the broker matters). Furthermore, the paper offers valuable methodological insights about using public data (TAQ) to document spikes in activity due to informational flows and data from various venues, including the development of an identification strategy using the geographic structure of latency. This contrasts with traditional

[11] The possibility of conflict of interest being experienced by the exchanges with respect to the pricing of data has been highlighted in various regulatory contexts being identified by the Securities and Exchange Commission (SEC), e.g., : https://www.sec.gov/rules/sro/nms/2020/34-88827.pdf, e.g., see discussion on page 9; https://www.sec.gov/rules/final/2020/34-90610.pdf (e.g., see section on page 200); and https://www.sec.gov/spotlight/equity-market-structure-roundtables/roundtable-market-data-market-access-102618-transcript.pdf (e.g., comments by SEC senior staff member David Shillman at pages 107-109).

approaches to measuring informational flows that focus upon the time series structure of returns or prices.

Additionally, this sheds light on the broader importance of HFT (high-frequency trading) in the equity markets. Our paper documents the race to execute trades and update quotes in response to the publication of off-exchange (dark) trades. In the broader context of information races, Aquilina, Budish and O'Neill (2020) use message data to calculate the value of a speed advantage. They identify races between different firms whereby both firms send orders within microseconds of each other, though they do not investigate the cause of these races. In contrast, we exploit geographic variation to allow us to pinpoint a specific information race: the race to execute trades and update quotes following the publication of dark trades, which we track across multiple exchanges. We are able to document bounds on the value of our race without using message data by documenting the negative realized spreads earned by traders in this race. We document $65 billion per year in trade responses and $775 billion per year in quote responses. With negative realized spreads of roughly half a basis point, this suggests a potential data value of $32 to $425 million per year, with the lower bound coming from the trade volume, and the upper bound coming from the trade and quote volume together.

II. Economics of Off-Exchange Trade Reporting

A. Trade Reporting Facilities

Before discussing the economics of trade reporting and data sales, we offer a brief technical summary of how trades are publicly disseminated. With exchange trades, when a trade occurs, all market participants are notified at the same time.[12] In contrast, off-exchange trades notify at least one party to the trade, but report the information to the public market in a lengthier two-stage process, which will be the focus of this section. First, the off-exchange trade is reported to a Trade Reporting Facility (TRF). Second, the TRF in turn reports the off-exchange trade to a Securities Information Processor (SIP), which broadcasts the trade to market participants.

There are currently three Trade Reporting Facilities (TRFs): one run by NYSE out of its Mahwah, NJ data center and two run by NASDAQ, with one in its Carteret, NJ data center and one in a Chicago, IL data center. In any off-exchange trade in an exchange-listed stock, one FINRA member broker has an obligation to report the trade to a TRF. Market participants have free choice

[12] For each exchange, the lowest-latency data feeds inform market participants of trades at the same time that any parties to the trade are notified.

over which TRF they report to: any NMS security from any tape can be reported to any of the three TRFs.

In turn, the TRF reports the off-exchange trade to the Securities Information Processor (SIP). These SIPs disseminate trades to general market participants, along with performing important regulatory functions such as calculating the National Best Bid and Offer. The complete broker to TRF to SIP process is depicted in Figure 1. There are two SIPs: The Consolidated Tape Association (CTA) SIP and the Unlisted Trading Privileges (UTP) SIP. The CTA SIP exclusively serves all Tape A and Tape B securities, while the UTP SIP exclusively serves all Tape C securities.[13]

Figure 1 provides a detailed overview of the reporting of data from off-exchange trading. A broker-dealer reports trades to one of several trade reporting facilities (which it chooses) and pays a reporting fee. These facilities report to the Securities Information Processors (SIPs), which sells and broadcasts their data to subscribers. Some of the SIP profits are rebated back to the trade reporting facility, which in turn rebates to the relevant broker-dealer.

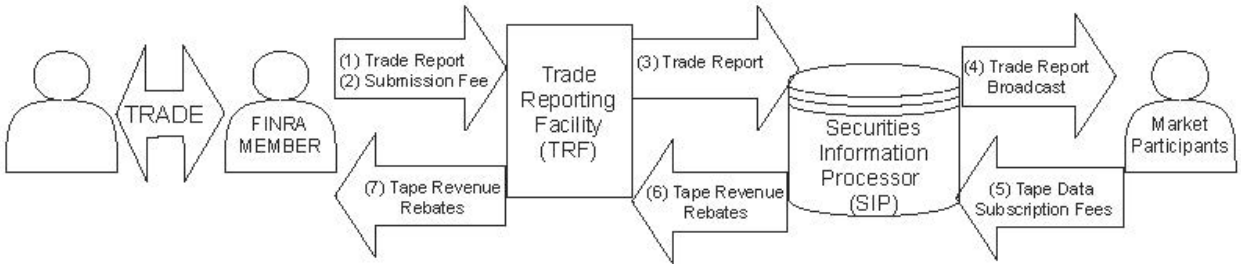
Figure 1. Trade Reporting Process for Off-Exchange Trades. When an off-exchange trade occurs, a FINRA member with the trade reporting obligation must report the trade to a Trade Reporting Facility (1), which charges a submission fee (2). The Trade Reporting Facility sends the trade to the Securities Information Processor (3). The Securities Information Processor (SIP) then broadcasts the trade (4). Market participants who subscribe to the SIP pay data subscription fees (5). The SIP earns a profit: these fees are rebated back to exchanges and trade reporting facilities according to a formula set by Reg NMS (6). The regulatory structure allows Trade Reporting Facilities to choose to rebate back some of the tape revenue that they receive to their members (7).

Per year, (5) is $400 million[14]. (6) is around $70 million. For (7), the rebates are returned to the TRF members in the form of securities transaction credits which are established by the TRF Business Member. As of January 1, 2021, traders with greater than 2% market share qualify for the largest tier, and would earn $1.8 million per year for each percentage of market volume their trades comprise[15] (7). Their annual trade submission fees to the TRF are capped at a payment of $360,000 per year (2). For smaller traders, the revenue sharing goes down, and the reporting fee per trade goes up.

[13] Tape A Securities are those listed by the NYSE. Tape B Securities are those listed on markets other than the NYSE and NASDAQ. Tape C Securities are those listed by NASDAQ. While a security may be traded on any exchange regardless of where it is listed, the trade report and compilation of the National Best Bid or Offer must be done by the SIP designated for a specific tape.

[14] The CTA and UTP SIP report tape rebates on a quarterly basis. https://www.utpplan.com/DOC/UTP_Revenue_Disclosure_Q32020.pdf and https://www.ctaplan.com/publicdocs/ctaplan/CTA_Quarterly_Revenue_Disclosure_3Q2020.pdf

[15] FINRA 7610A and 7610B record the rebate levels set by the NASDAQ TRF and NYSE TRF, respectively. For FINRA members in the top reporting tiers, NASDAQ shares 98% of attributable revenue while NYSE shares 100% of attributable revenue. "Attributable revenue" refers to the SIP Tape revenue rebates set forth by Reg NMS's "Revenue Allocation Formula", which allocates half of SIP profit to quotes, and half to trades. Of the SIP profit allocated to trades, it is paid out to each exchange or TRF according to the square root of exchange or venue nominal volume share of the square root of all nominal volume, adjusted for the size of the trade.

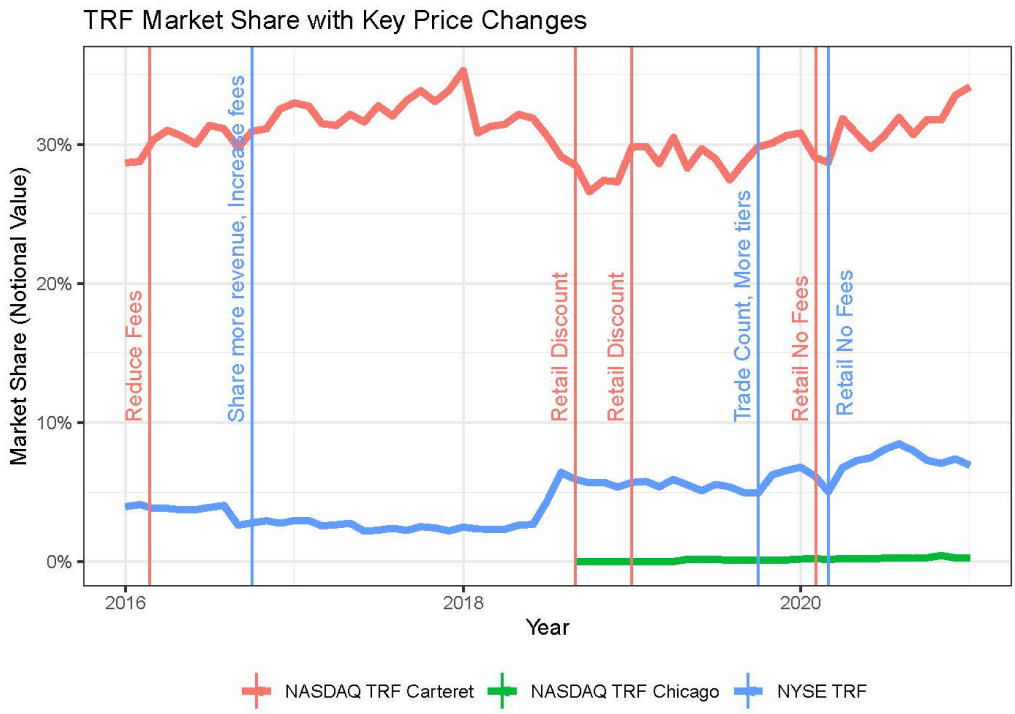


SIPs charge for the data they distribute. In 2019, the combined SIP income from data, net of administrative expenses, was $389 million. This combined pool of income is rebated to the exchanges and TRFs according to a formula set by Regulation NMS. Under the Revenue Allocation Rule,[16] the formula for rebates is set by Regulation NMS with 50% of the $389 million rebate pool allocated to exchanges posting NBBO quotes, and 50% of the $389 million rebate pool allocated to trades. These trade rebates are, in turn, allocated to the SRO reporting the trade based upon a formula adopted by plan participants pursuant to Regulation NMS. Figure 2 plots the market share of the major TRFs; with TRFs reporting 40% of all trade volume, they received $65 million in tape revenue rebates in 2019.[17]

Figure 2. Market Share of Trade Reporting Facilities. This chart documents TRF Market Share out of the total dollar volume of U.S. NMS securities. Key rule changes for each SIP are summarized by vertical lines. Rule changes either change the fees charged for trades, change the rebate formulas and tiers, or set up special discounts for firms which serve retail clients.

[16] CTA and UTP SIP revenue allocation formulas are set forth by the Revenue Allocation Formula: V.A.3 and Allocation Amendment V.B.1 of Regulation NMS (17 CFR Parts 200, 201, 230, 240, 242, 249, and 270. See page 37610 of https://www.sec.gov/rules/final/34-51808fr.pdf. See also https://www.utpplan.com/DOC/Revenue_Allocation_Formula.pdf.

[17] As noted with prior links to the CTA and UTP plans, the CTA and UTP SIP report tape rebates on a quarterly basis.



Like exchanges, Trade Reporting Facilities receive these SIP rebates and have determined, by rule, to pass some of the revenue back to firms submitting the trade reports. The parties running the TRFs set their own formulas, with SEC approval, for what portion of their SIP revenue that they pass back, with all facilities using a tier-based system.[18] For traders in the largest tier (with a market share greater than 2%), the NASDAQ TRF passes through 98% of attributable SIP revenue, and the NYSE TRF passes through 100%. For the traders in the smallest tier (with a market share less than 0.1%), 0% of attributable revenue is passed through. Both TRFs also charge trade submission fees, subject to SEC approval, again based on a tiered system with traders in the largest tiers receiving volume discounts. Based on regulatory fee filings, the rebates are substantially

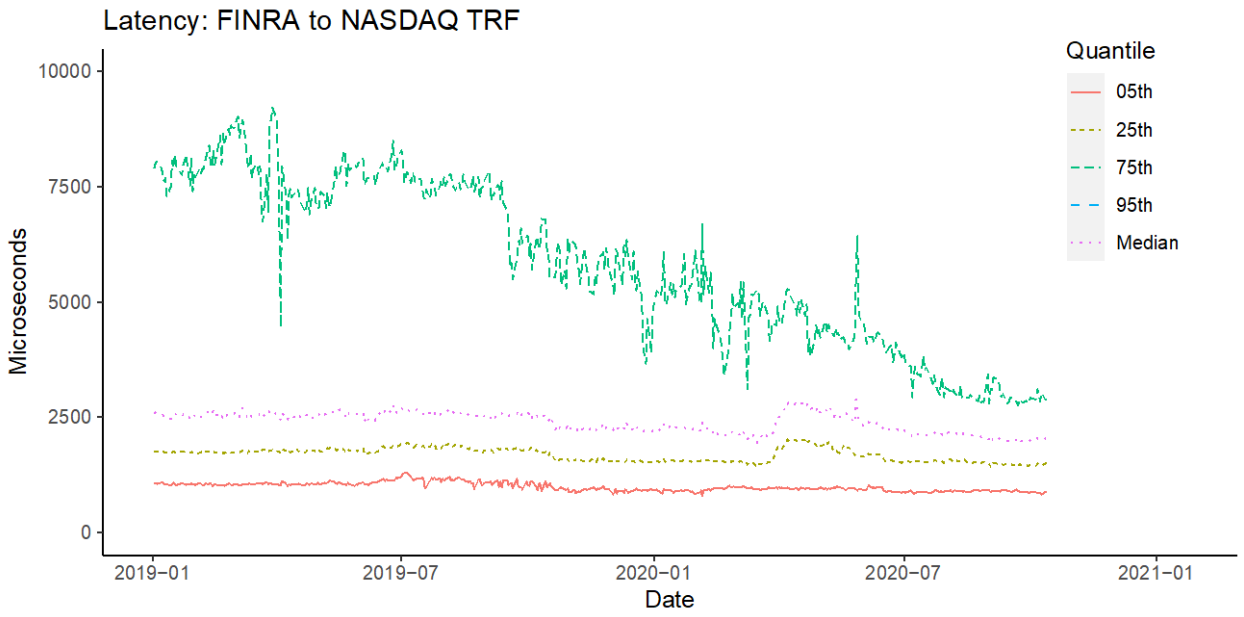
[18] The rules governing FINRA's TRF were established as part of the then NASD's separation from the NASDAQ at the time of the latter's exchange registration. See https://www.sec.gov/rules/sro/nasd/2006/34-54798.pdf. Prior to this time, NASDAQ was responsible for OTC trade reporting. As a result of this rule, NASD chose to maintain the existing reporting infrastructure and members were required to report OTC trades to the NASD/NASDAQ trade reporting facility. At the same time, NASD stated that it would be willing to enter into a similar agreement with any other exchange, and subsequently did so with the NYSE. See letter from Robert Glauber, NASD Chairman to the SEC, https://www.sec.gov/rules/sro/nasd/nasd2005087/nasd2005087-17.pdf.

larger than the fees charged.[19] Thus, the majority of TRF revenue comes from SIP rebates, and the majority of these SIP rebates are passed through to the reporting firms.

The latencies involved in the reporting of off-exchange trades are considerable. Figure 3 presents the time it takes for a trade report to reach the TRF from the FINRA member who reported it, while Figure 4 depicts the second stage of the journey, when the trade report must travel from the TRF to the SIP for broadcast. The combined trip can take anywhere from 1,000 to 10,000 microseconds. To put this into perspective, at the time an off-exchange trade occurs, a participant could send an order to trade from anywhere in New Jersey, have that order execute, have market participants notified of the trade, and then send another signal across the state of New Jersey before the off-exchange trade is published. In Section III we demonstrate that in spite of these delays, informative value of the trade remains, with market participants reacting quickly and precisely when the off-exchange trade is published.

Figure 3. FINRA Member to TRF Latency. In any off-exchange trade, one party to the trade will have an obligation report the transaction to a Trade Reporting Facility (TRF). This graph depicts several quantiles of the daily measurements of the time between the FINRA member's timestamp and the timestamp of the TRF (in this graph, restricted to be only the NASDAQ TRF). The FINRA member timestamp is only recorded to the nearest 1,000 microseconds. Note that for the typical day in the sample, the 95th percent quantile of latencies measured for that day is around 200,000 microseconds (i.e. 1/4th of one second), and is therefore far off the scale of this chart.

[19] In 2019, UTP and CTA SIPs report rebating a combined $10 million to the NYSE TRF and $24 million to the NASDAQ TRF. Per-participant trade reporting fees are given by 7620A and 7620 B, with monthly per-participant reporting fees capped at $30,000 per month for the NYSE TRF and $10,000 per month per tape for the NASDAQ TRF. Both rebates and fees are set in a tiered system; while an exact comparison of the fees and rebates would require knowledge of how many firms are in each tier, under the assumption that most tiers are not empty (an assumption supported by SEC release 34-88324, in which the NYSE TRF reports the number of firms in each tier, and SEC 34-84901, in which NASDAQ mentions the specific number of firms which would benefit from changes in retail trade reporting), overall rebates are much larger than the fees.

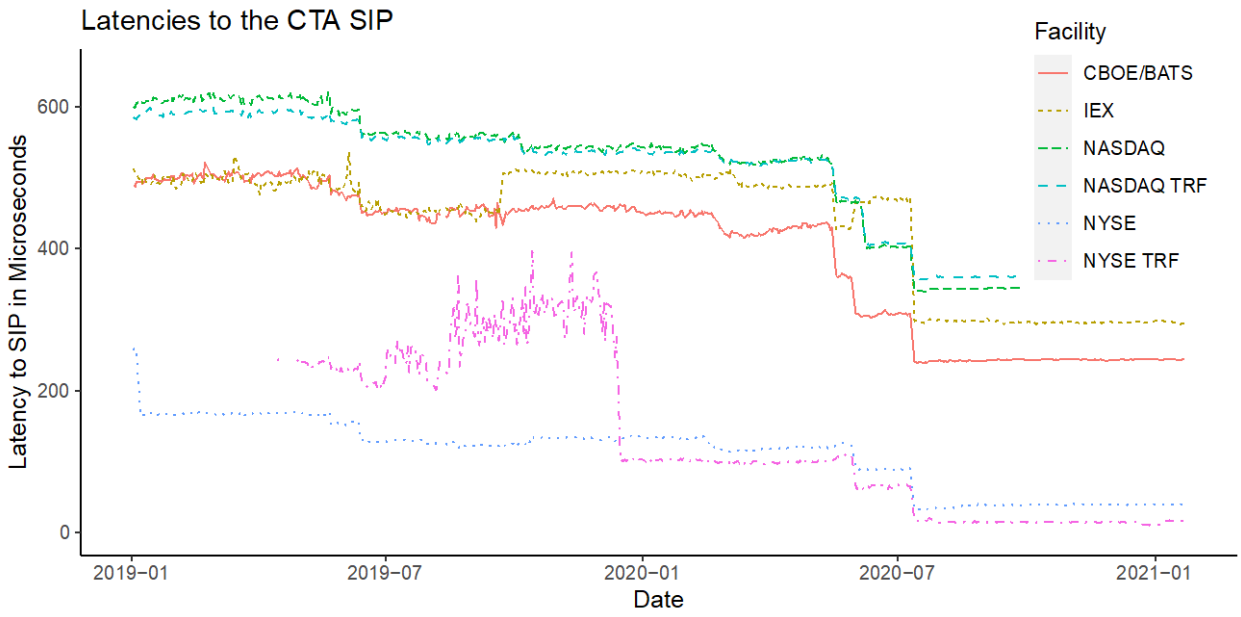


B. Economics of Data Pricing

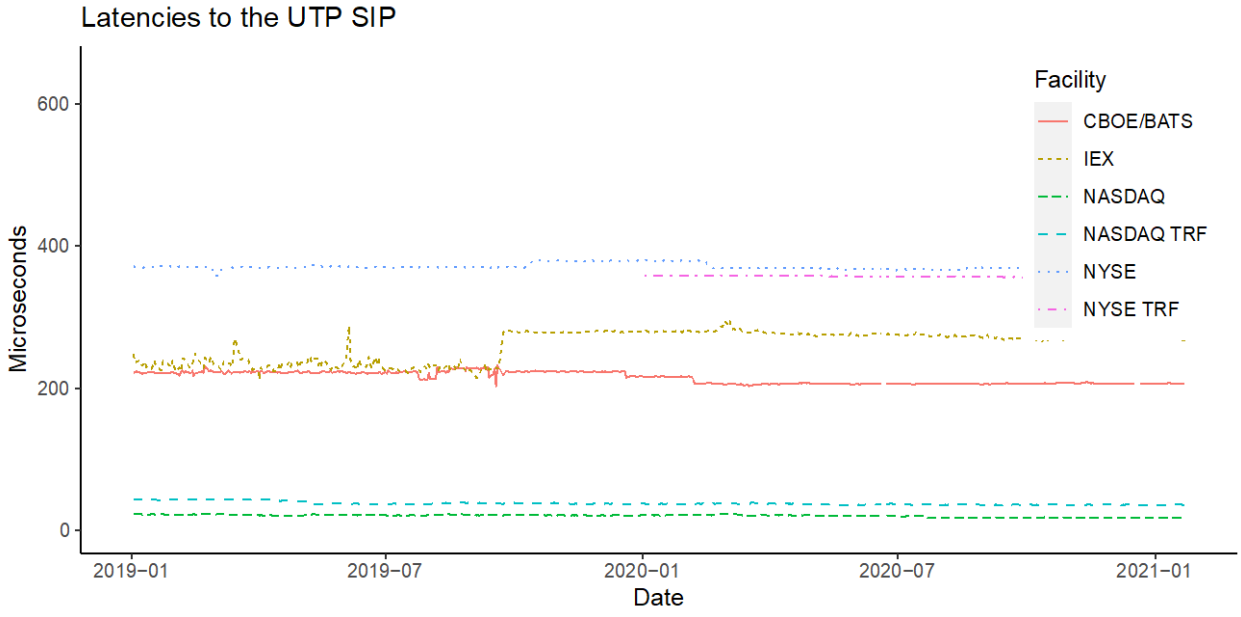
Information and data are central to our trading process. They provide value to market participants by allowing them to enhance their trading tactics. They are consequently a source of potential revenue to entities that can generate and offer the data, and the price of such data is an equilibrium estimate of the trade value of the information to other market participants. Meanwhile, the National Market System for trading equity is organized around transparency, creating a post-trade reporting obligation for trades due to the positive externalities of publication. In the case of exchange-based trades the reporting requirement is fulfilled by the exchange, while off-exchange trade reporting is undertaken by the trade reporting facility selected by the broker-dealer or by the ATS reporting the trade. The broker-dealer pays a modest fee to the trading reporting facility in order to be able to fulfill his obligation and can choose which reporting facility to use—currently among two operated by NASDAQ and one operated by the New York Stock Exchange (NYSE).

Figure 4. Latencies in SIP Reporting. All exchanges and TRFs send their trade reports to the SIP. The CTA SIP receives and processes trades in Tape A and B securities, while the UTP SIP receives and processes trades in Tape C securities. These figures plot the median daily latency, measured as the difference between the exchange or TRF timestamp and the SIP timestamp. This measurement of latency will include both the time for the signal to travel the geographic distance across New Jersey as well as server computation time in sending, receiving, or processing the trade. The server computation time is considerably higher at the CTA SIP compared to the UTP SIP, though the difference decreases throughout the sample period.

Panel A: Latencies from facilities to the CTA SIP



Panel B: Latencies from facilities to the UTP SIP



This leaves open basic questions about how data should be priced, especially in light of the classic perspective in economics, which regards information and data as a public good, suggesting that market participants should not be excluded from information, i.e., that data pricing should be close to zero. Because of the separability between the order routing decision and the TRF routing decision, the TRF or the trading platform cannot be punished by routing order flow elsewhere due

to high data prices (unlike exchange order routing); on the other hand, much of the SIP revenue is being rebated. While firms are required by FINRA to be connected to at least two TRFs,[20] there still is considerable potential competition for the TRF reports because of the presence of multiple facilities through which to report (two owned by NASDAQ and one by the NYSE).

The interaction between trading and market data in the exchange trading context is highlighted by Spatt (2021). Due to the value of market data (and connectivity services) exchanges compete vigorously to attract liquidity and provide trading services, which in turn enhances the value of the data and connectivity being provided by the exchange to all market participants. Spatt (2021) argues that profit-maximization by the exchanges leads to cross-subsidization of trading (through low or even negative net fees) as exchanges seek activity to enhance the value of the data (and connectivity) that they can offer. Despite the cross-subsidization, in recent years about 40% of trading in NMS equities has been off-exchange.[21] The last year has seen even greater concentration of trading occurring off-exchange.[22]

There is an underlying perception that data emanating from the exchanges is relatively more valuable than trade reports from other platforms—both because off-exchange trading is more retail oriented[23] and because quote and order book information is not present in the off-exchange context as the TRFs never report quotes or order book information, only trades. This could be a reflection of both the anticipated limited value of quotes and orders, if it were readily available, and the design of the various off-exchange trading mechanisms with respect to quotes and orders. In the exchange context there is greater scope for valuable information to be generated because the off-exchange mechanism focuses upon (less informed) retail orders and does not produce quote and order information. However, this does not suggest that there is no value to off-exchange trade

[20] In 2016, FINRA published a Regulatory Notice entitled "OTC Equity Trading and Reporting in the Event of System Issues," which required firms to establish and maintain connection to a secondary TRF reporting facility in case of a disruption at the firm's primary reporting facility. See https://www.finra.org/sites/default/files/notice_doc_file_ref/Trade-Reporting-Notice-012016.pdf.

[21] On the issue of cross-subsidization to enhance the value of exchange data, it also is interesting to note that in contrast, IEX (Investors Exchange) does not sell its data and consequently charges substantially higher trading fees. Arguably, then trading is subsidizing data, which is the opposite of the pattern implemented by the incumbent exchanges, reflecting a differentiating strategy by IEX.

[22] TRF notional volume share has increased from 35% of the market to 40%, while the volume share has increased from 40% to around 50%. Values calculated from TAQ and CBOE data.

[23] For example, many of the most recent Trade Reporting Facility rule filings are about special pricing for retail brokers.

reports, but just relatively less demand than for exchange data—and potentially less focus on accessing and utilizing the off-exchange data by market participants.

Nevertheless, our empirical findings suggest that there is a significant response to the publication of TRF trade reports by the Securities Information Processor (SIP). Surprisingly, we observe no change in the latency pattern when NYSE TRF Trades are included in some proprietary NYSE data feeds. Both before and after their inclusion, the latency pattern around the TRF trade reports remains the same: the spike in trade and quote volumes continues to align with the geographic latency from the SIP data center, and not the TRF data center. This poses a mystery: why are market participants reading TRF trades from SIP data, and not a faster proprietary feed?

Consistent with practice described by market participants, this suggests that the particular feeds are ones that are not typically utilized by high frequency trading firms. However, it also is surprising that the off-exchange trades are only included in the cheaper exchange feeds, and not in the expensive feeds marketed to HFT firms. Finally, in contrast, there is not a large response to the publication of exchange trades by the SIP, as many investors learn about these exchange trades from proprietary data feeds. In summary, the response to the publication of exchange trades appears to be linked to the reporting from the proprietary data feeds rather than from the SIP, while the response to the publication of off-exchange trading does not occur until its publication through the SIP.

It's interesting to reflect further on the changes in the structure of trade reporting and its pricing. In the aftermath of the adoption of Regulation SCI (Systems Compliance Integrity) and as a result of FINRA's Trade Reporting Notice, broker- dealers were required to have direct access at least two alternative trading reporting mechanisms for off-exchange trades to guard against the possibility of system failure. Previously, the NASDAQ and NYSE each offered a single reporting mechanism, but NASDAQ was by far the dominant TRF, with over 90% share of the off-exchange reporting. NASDAQ developed a second TRF in Chicago in response to FINRA's Trade Reporting Notice together with the NASDAQ TRF in Carteret, brokers would be able to satisfy their dual-connectivity requirement by connecting to the two distinct NASDAQ TRFs (allowing participants to use its reporting service to have available a second NASDAQ facility to be compliant with FINRA requirements.

However, in the aftermath of the requirement to maintain a primary and secondary TRF reporting connection, the market shares of trade reporting moved somewhat away from the

dominant player (NASDAQ) and towards the secondary player (NYSE). The need for broker-dealers to maintain direct access to multiple trade reporting facilities led to greater competition in pricing (in the form of higher rebates of SIP revenues) because broker-dealers were required to incur the fixed cost of access to a second facility anyway and some shift of reporting volume to the NYSE facility resulted. It also is interesting to note that the reporting facilities use pricing tiers, so that the larger brokers obtain higher rebates (and smaller brokers have incentives to report through larger brokers if that would enhance their effective rebate).[24] In any case, the use of the pricing tiers should encourage individual brokers to route all of their activity through a single TRF or be at the maximums of the respective pricing. NASDAQ's SIP rebate went from 50% to 98% for traders in the top tier, reflecting the much greater competition that emerged in the aftermath of the dual-reporting requirement and the need to incur fixed costs to be able to report to a second facility (leading to back-up by the NYSE TRF). By broker-dealers using the NYSE TRF as a secondary facility rather than a second NASDAQ TRF limits the monopoly of the TRFs in a way that would not arise when subscribing to two NASDAQ TRFs. This would encourage greater price competition between the trade-reporting facility of the NYSE and one of the two NASDAQ facilities, because having immediate access to TRFs operated by both exchanges would enhance the ability to easily respond to price changes from the TRF operated by the other exchange.

This is an interesting and perhaps unexpected consequence of FINRA's requirement for broker-dealers to maintain two TRF connections. Though nowhere near as complex as the pricing tiers for equity orders that are used to encourage liquidity in exchange trading (see Spatt (2021)), the price discrimination (toward larger players) does have significant elements in common, such as encouraging the routing (of reporting) of marginal activity to large firms and encouraging smaller brokers to consolidate their trade reports with larger entities.

The presence of cross subsidization, such as the extent of rebates and net costs on alternative reporting regimes, will influence the reporters' choice of trade reporting facility. The TRF choice is separable from that of an off-exchange platform to execute an order as the reporting facility can be determined independently of execution (unlike the choice of exchange, where the reporting is bundled with execution on a particular exchange).

[24] There would not be an advantage to the smaller broker reporting to the TRF through a larger broker, if the smaller broker were receiving the top tier (maximum) rebate.

Brokers decide how and where to execute orders, possessing some flexibility to execute on various platforms—both on and off exchanges. These execution decisions are influenced by the desire to obtain high quality executions for their customers, though constrained in the exchange space by Regulation NMS and constrained on an overall basis by "Best Execution" responsibilities. The standards for "Best Execution" require a broker to exercise reasonable care to execute a customer's order in a way to obtain the most advantageous terms for the customer. This, in essence, emphasizes the importance of the underlying process and provide some discretion to brokers. Our framework suggests balancing and integrating it against fees, rebates, and data prices (costs and trading activity executed by the broker-dealer should influence the data being acquired) and that "Best Execution" could have different meaning for various participants based upon such considerations as latency, internalized order flow and the data available to the firms. For those orders that are executed "off-exchange" the broker has a routing decision with respect to trade reporting (which trade reporting facility to utilize?), which is separate from the order routing decision.

It should be noted that making the TRF data free would not lead to a second-best solution, since the exchanges are able to cross-subsidize and attract trading at the margin and sell their data. Another important facet of the pricing of exchange data is the extent to which trade vs. quote data is more valuable and how each is priced.

The pricing of proprietary data and competition between proprietary data and the SIP has become a widely debated aspect of the plumbing of our regulatory system for equity market trading. The SIP data has been extensively criticized because it has not benefited from the same enhancements as proprietary data in terms of speed and detail.[25] These perspectives underlie some of the recent reform and modernization of the structure of the SIP that was approved by the SEC. This criticism is associated in part with the agency conflict between the governance of the SIP and the management of various proprietary data. The recent reforms adopted by the SEC (https://www.sec.gov/rules/final/2020/34-90610.pdf) were designed to expand the content of NMS market data (including some information about odd lots, depth of book and orders

[25] It also has been criticized because of the extent of control by the exchanges in the governance and pricing of the SIP (the relative lack of improvement in the SIP may be viewed as a consequence of the structure of its governance).

participating in auctions) and encourage competing consolidators, rather than an exclusive SIP largely controlled by the incumbent stock exchanges.

As described by the SEC, the SIP overhaul is motivated in part as an attempt to solve the agency conflict over the control of the SIPs, but it doesn't address the potential for agency conflict between the SIP and off-exchange trade reporting. In particular, it is a curious accident of history that the reporting facilities for the trades that compete with the exchanges, i.e., trade reporting facilities for off-exchange trading, are operated by the exchanges. Although there may be some system-wide savings for the reporting utility not to be built multiple times, exchanges do compete with brokers for execution services.

On the other hand, the practices of the Trade Reporting Facilities may not be especially aggressive in order for the exchanges to avoid exploiting the conflict of interest. For example, perhaps this explains why the trades from the Trade Reporting Facilities are not provided in the proprietary feeds targeted for HFT firms and instead only included in those oriented to a broader audience. An alternative interpretation is that this helps protect the value of the proprietary data of the exchanges intended for HFT investors.

III. Identifying Market Reaction to TRF Trades

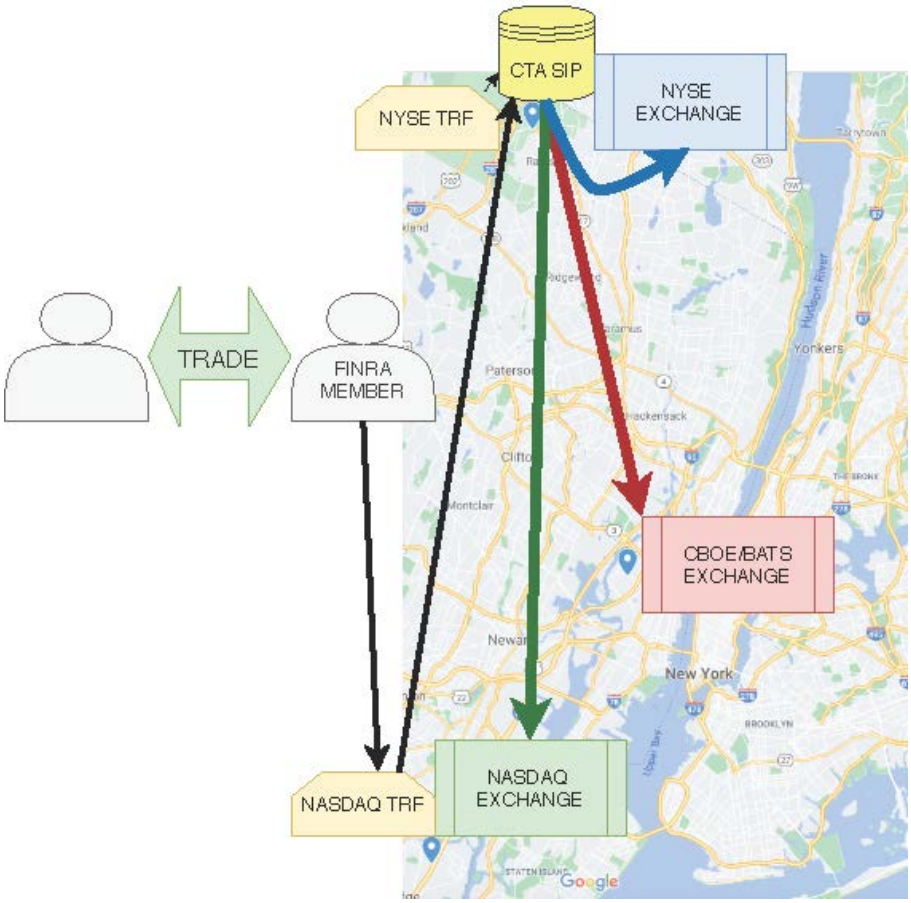
A. Geographic Variation in Trading

In the United States, equity trading and reporting is centered on the New Jersey Equity Triangle. Each vertex of this triangle is formed by a data center for one of three major exchanges: NYSE, NASDAQ, and CBOE. We exploit this geographic variation to identify the reaction of traders to the publication of off-exchange trade data. The reaction of market participants is clearly identifiable, with the publication of TRF trade reports leading to a sharp, sudden increase in both trading and quoting activity. This rapid increase in activity precisely aligns with known exchange-to-SIP latencies, with the time that the SIP signal first arrives at the exchange matching the point at which exchange trading and quoting activity sharply increases. Results are consistent across each unique exchange- to-SIP pathway; from the two New Jersey TRFs, two SIPs, and three exchanges, there are twelve independent pathways for off-exchange trade reports to reach market participants.

When an off-exchange trade is published by the SIP, there is a unique geographic pathway between each exchange and the SIP. The exchange-to-SIP latencies are depicted in Figure 5 for the CTA SIP, and in Figure 6 for the UTP SIP. If a signal is released by the CTA SIP in northern New Jersey, the signal arrives first at the NYSE data center in northern New Jersey, second at the CBOE/BATS data center in southeastern New Jersey, and lastly at the NASDAQ data center in southern New Jersey. Traders who react to SIP information would do so at specific times at each location, upon the arrival of the information. These reaction times line up with the known, relatively constant geographic latencies between the SIP and major exchanges.

We measure trading and quoting activity at each of the exchanges around the time that the TRF trades are published by the SIP. The market reaction to off-exchange trade reports is strikingly sharp, and consistent with geographic latencies. Figure 7 plots quoting activity around the SIP publication of TRF reports in Tape A securities (which are published by the CTA SIP). Following the publication of a trade, there is a rapid and brief increase in quoting activity occurring sequentially at each of the major exchanges. These sharp increases in quoting activity align with the geographic latencies given by Figure 5: 260 microseconds for NYSE, 380 microseconds for BATS, and 500 microseconds for NASDAQ. Moreover, these spikes in quotation volume are sharp: there is far less activity outside these exact points.

Figure 5. NJ Market Centers for Tape A Securities. The CTA SIP in Mahwah, NJ processes trades and quotes for Tape A Securities. Brokers have a choice to which TRF they report; both the NYSE TRF and NASDAQ TRF report to the CTA SIP for Tape A Securities. The SIP then broadcasts the report of these trades to market participants. Median exchange-to-CTA latencies for exchange trades in TAQ are 260 microseconds for NYSE, 380 microseconds for BATS, and 500 microseconds for NASDAQ.



B. Distinct Trading Facilities

We take advantage of the two distinct TRFs to confirm that traders are reacting to the SIP publication of the TRF trade, and not some other simultaneous event. For example, suppose a broker who trades off-exchange also places a simultaneous trade on an exchange. We can confirm that participants are reacting to the SIP publication of the TRF trade, and not a simultaneous action, by measuring latency responses for each TRF separately. As detailed in Section II. A, brokers have free choice over which facility to which they report trades. For securities processed by the CTA SIP in northern New Jersey, the TRF-to-SIP pathway is far shorter for the NYSE TRF, which is located in northern New Jersey than for the NASDAQ TRF, which is located in southern NJ. This

creates two distinct paths for the information to reach market participants, with the SIP timestamp serving as a shared reference point.

Figure 6. NJ Market Centers for Tape C Securities. The UTP SIP in Carteret, NJ processes trades and quotes for Tape C Securities. Brokers have a choice over which TRF they report to, with both TRFs reporting to the UTP SIP for Tape C Securities. Median exchange-to-UTP latencies for exchange trades in TAQ are 370 microseconds for NYSE, 220 microseconds for BATS, and 23 microseconds for NASDAQ.

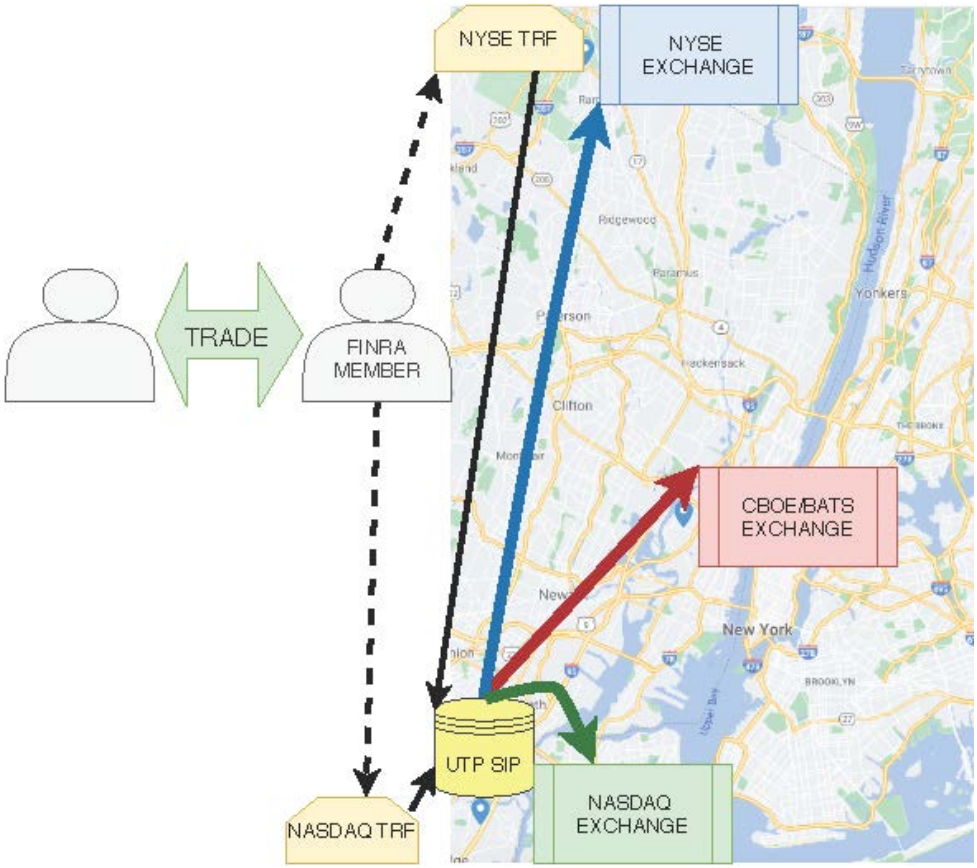
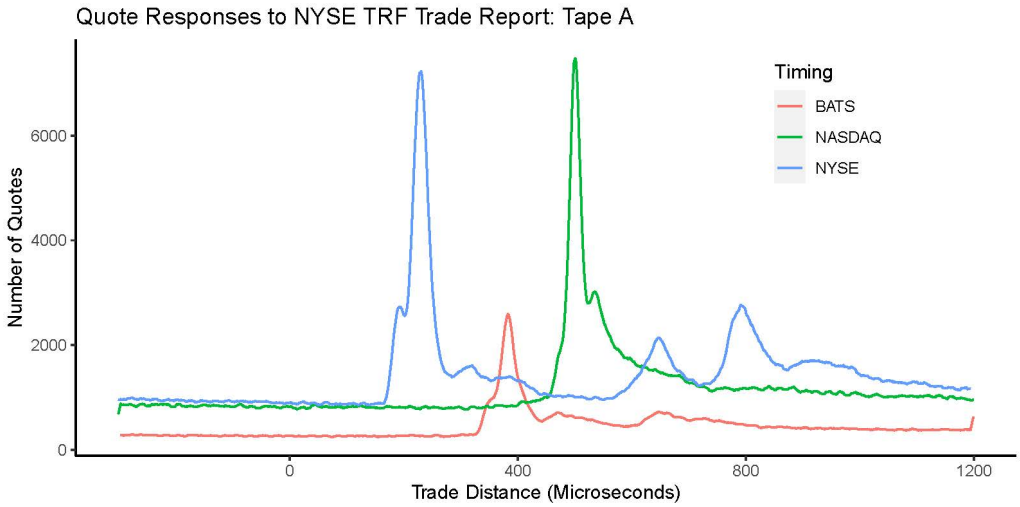
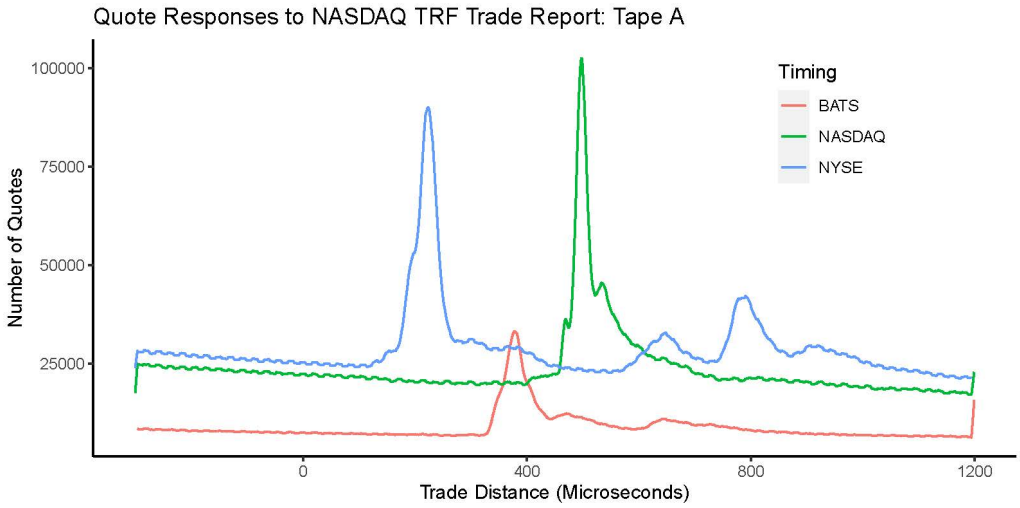


Figure 7. Quote Patterns around Tape A TRF Report. Depiction of the quoting pattern around the publication of a Tape A TRF trade report for stocks in the data sample. Time zero denotes the SIP publication time of the TRF trade report. The x-axis measures the time between the SIP TRF timestamp, and the exchange timestamp of quotes. The y-axis denotes the total volume of quotes occurring at each possible offset. There are sharp increases in quoting activity around 260 microseconds for NYSE, 380 microseconds for BATS, and 500 microseconds for NASDAQ.

Panel A: Trade Reports from the NYSE TRF



Panel B: Trade Reports from the NASDAQ TRF

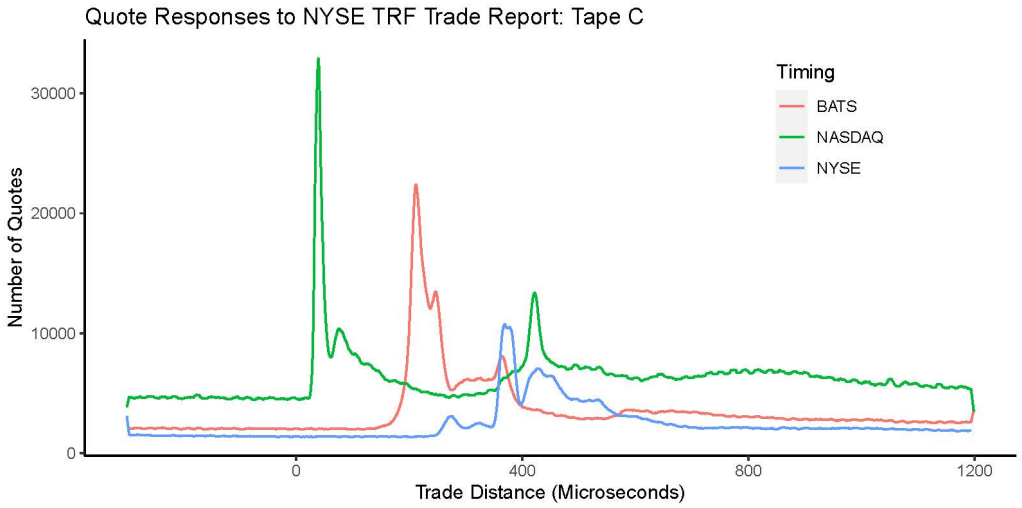


The reaction times to the SIP publication of a TRF trade in a Tape A security is the same whether the trade is reported through the NYSE TRF or NASDAQ TRF. Figure 7 separately plots the pattern of quoting activity for the NYSE TRF in Panel A and the NASDAQ TRF in Panel B. The figures are nearly identical, with spikes in quoting volume showing up at the exact same latencies from the SIP publication of TRF trades. If a broker places a simultaneous on-exchange and off-exchange trade, the trades occur in different data centers, and have distinct latencies to

reach all market participants. The off-exchange trade has two possible paths depending on which TRF is selected, but the SIP timestamp provides a common reference point for trades. The consistent reaction to the SIP timestamp across the two TRFs presented in Figure 7 confirms that market participants are responding to the TRF trade publication, and not an alternative event.

Figure 8. Quote Patterns around Tape C TRF Report. Depiction of the quoting pattern around the publication of a Tape C TRF trade report for stocks in the data sample. Time zero denotes the SIP publication time of the TRF trade report. The x-axis measures the time between the SIP TRF timestamp, and the exchange timestamp of quotes. The y-axis denotes the total volume of quotes occurring at each possible offset. There are sharp increases in quoting activity around 40 microseconds for NASDAQ, 280 microseconds for BATS, and 380 microseconds for NYSE.

Panel A: Trade Reports from the NYSE TRF



Panel B: Trade Reports from the NASDAQ TRF

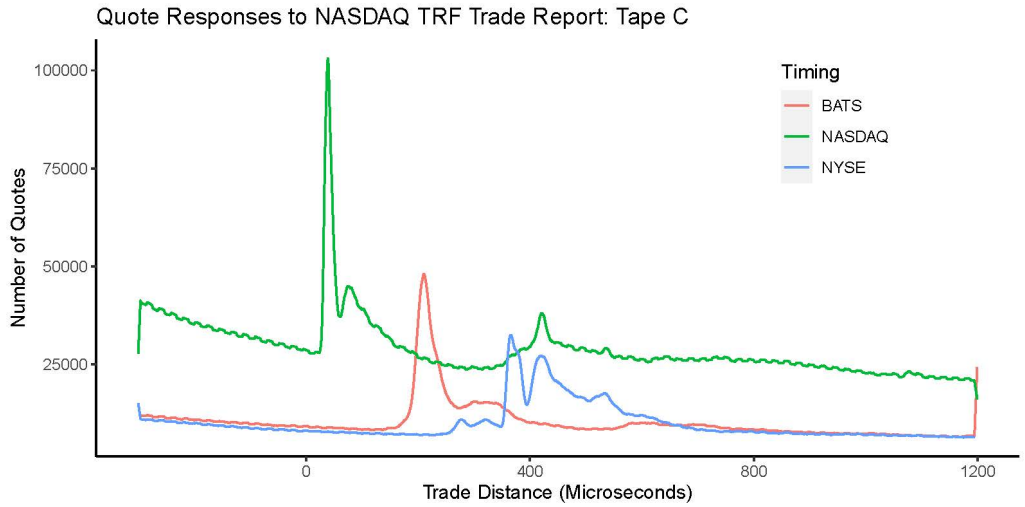
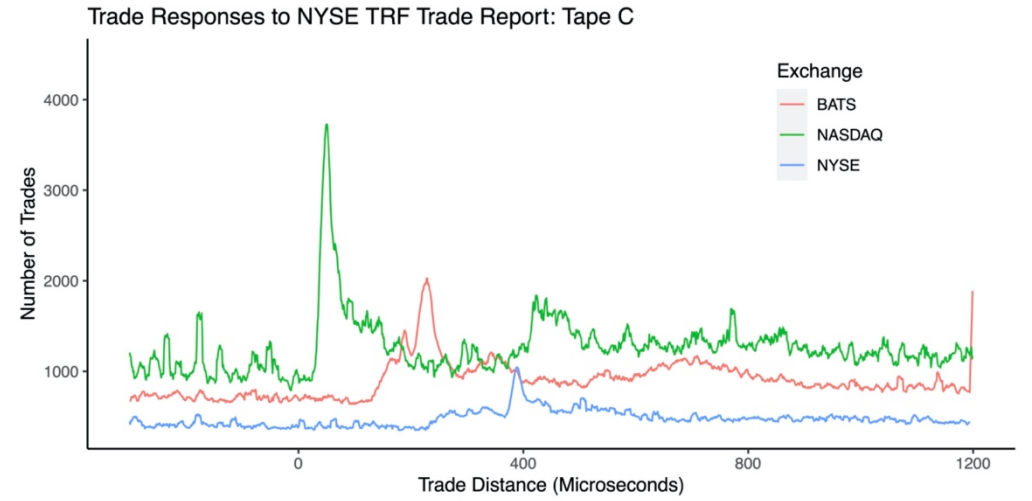
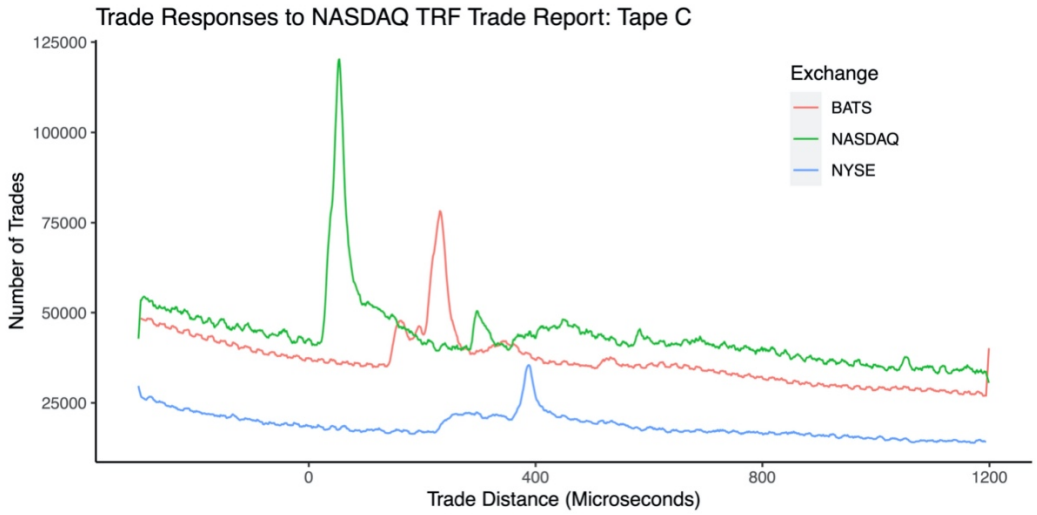


Figure 9. Trade Patterns around Tape C TRF Report. Depiction of the trading pattern around the publication of a Tape C TRF trade report for stocks in the data sample. Time zero denotes the SIP publication time of the TRF trade report. The x-axis measures the time between the SIP TRF timestamp, and the exchange timestamp of trades. The y-axis denotes the total volume of trades occurring at each possible offset. There are sharp increases in trading activity around 40 microseconds for NASDAQ, 280 microseconds for BATS, and 380 microseconds for NYSE.

Panel A: Trade Reports from the NYSE TRF





A second, distinct variation in latency pathways arises from variation in the Securities Information Processors. Trades in Tape C are processed by the UTP SIP, which is located in southern New Jersey. Figure 6 depicts the latency paths from the UTP SIP to each exchange, with the UTP SIP being closest to the NASDAQ exchange and the NYSE exchange being furthest from the SIP. In addition to a distinct geographic positioning, the UTP SIP also has a different average

processing time for trades, leading to shorter total latencies. Figure 8 plots quoting activity while Figure 9 plots trading activity around the SIP publication of TRF reports in Tape C securities (which are published by the UTP SIP). Following the publication of a TRF trade, there is a rapid and brief increase in trading and quoting activity occurring sequentially at each of the major exchanges. These spikes in trading and quoting activity are reversed in order relative to Tape A securities, as the first spike in trading now occurs at NASDAQ, the second spike at CBOE/BATS, and the last spike at NYSE. As before, the pattern in exchange activity is the same whether the trade report originated from the NASDAQ TRF or the NYSE TRF.

The combined variation of two distinct TRFs, two distinct SIPs, and three exchanges gives twelve unique pathways for an off-exchange trade report to reach market participants. This fixed set of different paths allow us to rule out alternative explanations for the patterns we observe. At the time of an exchange trade, many possible simultaneous events may occur. Brokers may place simultaneous trades on an exchange, or they may have placed their off-exchange trade in response to an event. But each of these events take a different path to market participants, and what we observe in the data is a rapid, sharp response to the SIP timestamp of off-exchange trades. For each SIP, we observe the same latency reactions to the SIP publication. By comparing Tape A and Tape C, we confirm that the reactions are precisely aligned to the SIP-to-exchange latency across each of the three major exchanges that form the New Jersey Equity Triangle.

C. Data

Our investigation of trading behavior around off-exchange trade reports relies on microsecond TAQ data. We analyze all trades and quotes from January 1, 2019 to Dec 31, 2020 for a sample of 300 stocks consisting of the 100 most-traded securities by volume from each of Tape A, Tape B, and Tape C. These trades were cleaned according to the techniques outlined in Holden and Jacobsen (2014). We obtain data on common daily liquidity measures, calculated from TAQ, from Conrad and Wahal (2020).

We measure trading or quoting activity on the exchanges against the publication time of off-exchange trades. For most of our analysis, we rely on the SIP timestamp of the off-exchange trade,

and compare it against the exchange timestamps of exchange trades.[26] In some analysis, we utilize the TRF timestamp of off-exchange trades, and compare it against exchange trading volume.

Comparing timestamps from different facilities does raise the possibility of clock drift. At microsecond timescales, clocks at different data centers may deviate from each other. As part of the SEC's National Market System Plan Governing the Consolidated Audit Trail, the exchange and SIP clocks are synchronized to within 100 microseconds of NIST time, though CAT NMS surveys report average maximum deviations from NIST of 36 microseconds or less. These deviations from NIST time add mean zero noise to our data observations. The market activity reaction spikes like those in Figure 8 will have a wider span of elevated quoting activity driven by this clock drift, along with variation in order processing at the SIP and exchange.

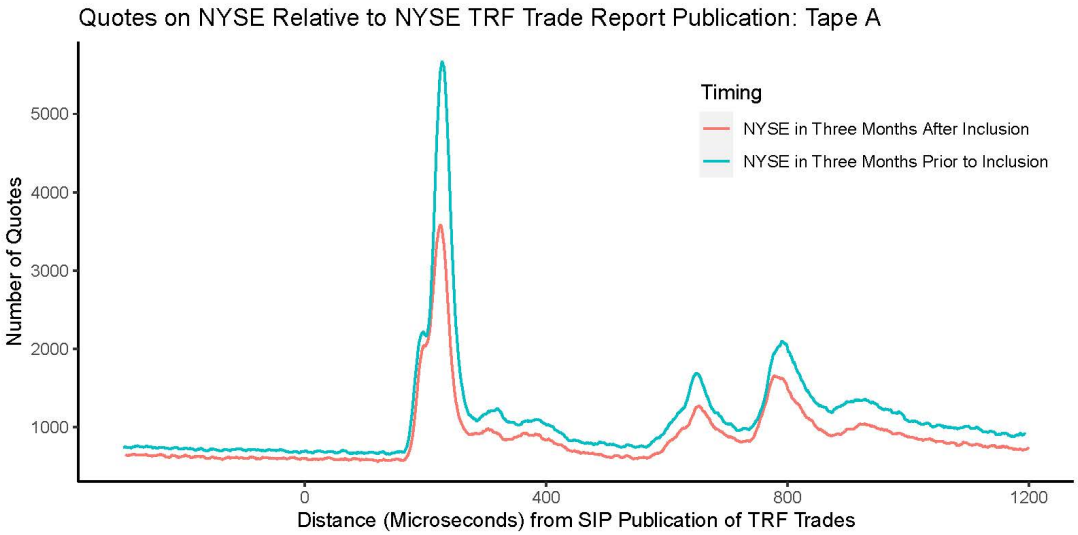
D. Proprietary Feed Inclusion

While Trade Reporting Facilities report trades to the SIPs, the TRF operating business partners have the option to include their off-exchange trades in a proprietary feed. Both NYSE and NASDAQ include TRF trades in their proprietary feeds. NASDAQ has done so for the duration of our sample period, but NYSE began including TRF trades on April 29th, 2019. Figure 10 analyzes the pattern of quoting activity on the NYSE Exchange before and after the NYSE TRF included TRF trades in proprietary feeds. There is no change in the pattern of activity, with the spike in response trades occurring at the same delay from the SIP publication timestamp. This suggests that the market participants placing these quotes read the information from the SIP broadcast, and not a proprietary feed.

Both NYSE and NASDAQ include their respective TRF data in only a select set of their proprietary feed offerings. Notably, neither NYSE's OpenBook nor NASDAQ's TotalView include TRF trades. Both of these feeds represent the premier offerings by each exchange, with full depth of book and the lowest latency available. The proprietary feeds which include TRF trade information are the cheaper feeds from each exchange, and are priced closer to the price of SIP data.

[26] We have also measured against the SIP timestamp of exchange trades. This pushes back the latencies, as it must include a round trip from both the SIP to exchange, and from the exchange back to the SIP. Due to the additional noise and complication from this second leg of the journey, we present results on the exchange timestamps.

Figure 10. TRF Inclusion in Proprietary Feed Data. The NYSE TRF began transmitting TRF trade information to proprietary data feeds on April 29th, 2019. Quote activity on the NYSE Exchange dramatically increases around 180 microseconds after the SIP timestamp of an off-exchange trade. This pattern of activity remains the same after NYSE incorporated TRF reports in select proprietary data feeds.



E. Timestamps Analysis

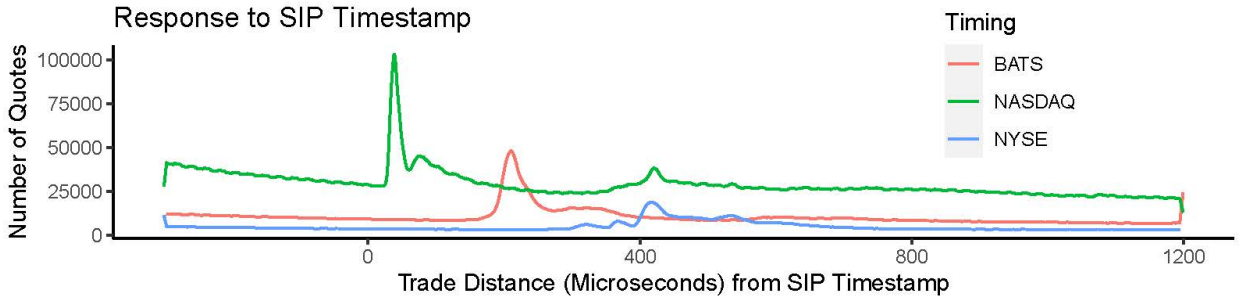
The reporting time, as well as the market response time, differs considerably between TRF trades and exchange trades. While TRF trades are followed by a dramatic increase in market activity in response to the SIP publication of trades, exchange trades show no reaction at all to SIP publication of trades. We document these differences in market reactions, as well as the differences in reporting delays between on-exchange and off-exchange trades.

Both the exchange and the TRF timestamp trades before sending them to the SIP.[27] For the TRF trades, there is no market reaction until the trade is published by the SIP. Once the TRF trade is published by the SIP, there is the sharp, dramatic increase in market activity which aligns with the geographic latencies from the SIP. Figure 11 plots this activity. The pattern from the SIP and TRF Timestamps are almost identical, with the only difference that the TRF pattern occurs at a further delay of around 20 microseconds. This reflects the time for the signal to travel from the TRF to the SIP for broadcast, as well as some processing time by the exchange.

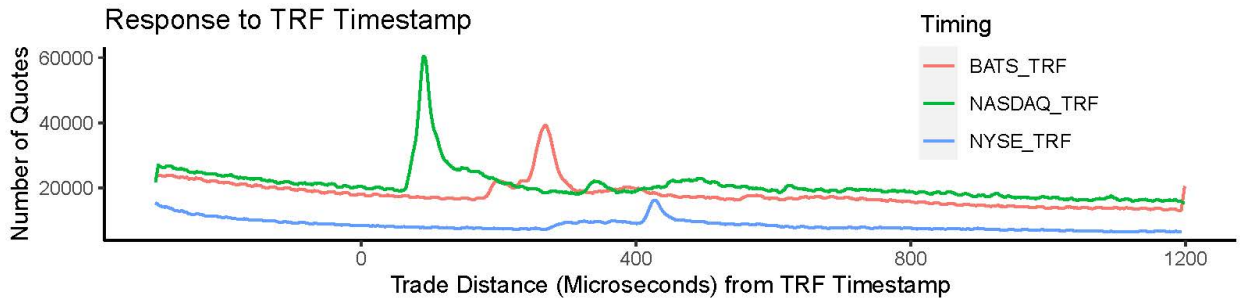
[27] TRF timestamps are provided only if the data is transmitted to a proprietary feed, as per the SIP technical specifications. Trades from the NASDAQ TRF have TRF timestamps throughout our sample, while the NYSE TRF trades have a TRF timestamp only from April 29th, 2019 onwards. Before this date, the NSYE TRF trades have a blank TRF timestamp field.

Figure 11. Comparison Between TRF Timestamp and SIP Timestamp. Off-exchange trades are timestamped by the Trade Reporting Facility and the SIP. Panel A measures quoting activity on the exchanges relative to the SIP timestamp of NASDAQ TRF trades for Tape C securities in the sample. Panel B measures quoting activity on the exchanges relative to the TRF timestamp of NASDAQ TRF trades for Tape C securities in the sample. Patterns of activity look similar, with the quote responses to the TRF timestamp occurring 20 to 40 microseconds later than the responses measured against the SIP timestamp.

Panel A: Exchange quoting activity measured against SIP Timestamp.



Panel B: Exchange quoting activity measured against NASDAQ TRF Timestamp.

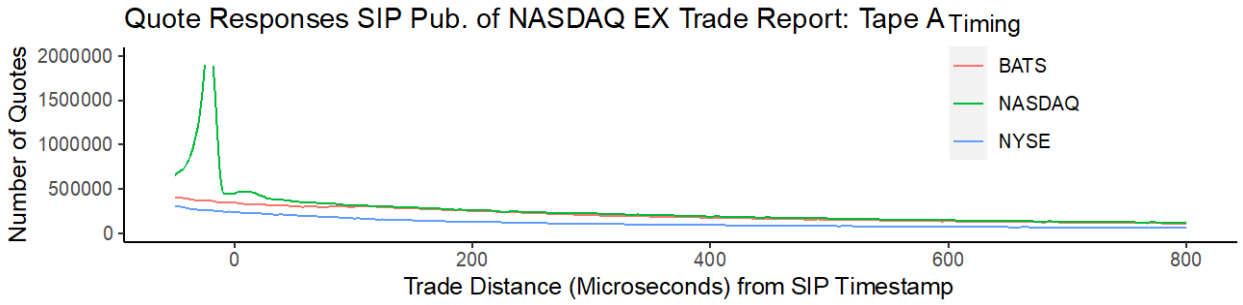


In contrast, the pattern in market activity around an exchange trade varies considerably between the exchange timestamp and the SIP timestamp. Figure 12 depicts this difference. Around the time of the exchange trade, there is a sudden, sharp increase in market activity on the exchange, which quickly drops off. At the time of the subsequent SIP timestamp, there is no reaction to the trade, with a constant level of market activity following the SIP broadcast of the trade.

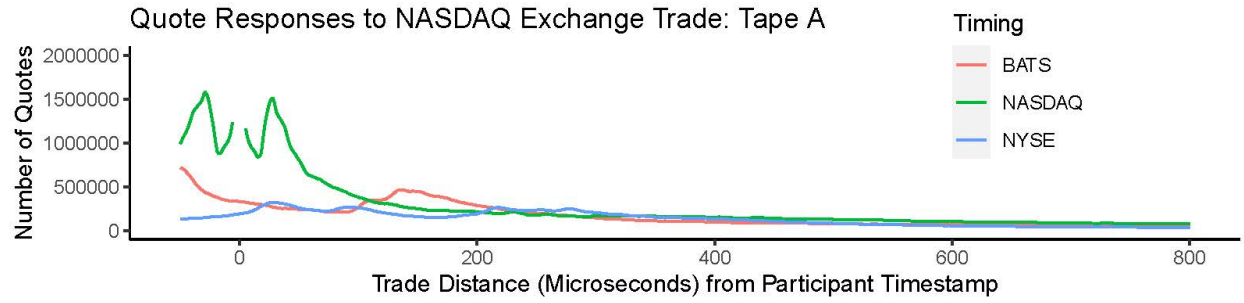
On-exchange trades are rapidly reported to the SIP, with Figure 4 depicting the typical latencies between the exchange execution and SIP timestamp of trades. For the UTP SIP (Tape C Securities), NASDAQ exchange trades are reported by the SIP within 20 microseconds. Reports from the NYSE exchange to the SIP take around 380 microseconds, reflecting the longer geographic distance which must be covered. For the CTA SIP, the same geographic patterns exist, but there is a longer processing time.

Figure 12. Comparison Between Exchange Timestamp and SIP Timestamp. Exchange trades are timestamped by the exchange and the SIP. Panel A measures quoting activity on the exchanges relative to the SIP timestamp of NASDAQ Exchange trades for Tape C securities in the sample. Panel B measures quoting activity on the exchanges relative to the exchange timestamp for NASDAQ exchange trades for Tape C securities in the sample. There is a sharp, dramatic increase in quoting activity around the exchange timestamp of NASDAQ exchange trades. Around the SIP timestamp, there is no change in market activity.

Panel A: Exchange quoting activity measured against SIP Timestamp.



Panel B: Exchange quoting activity measured against NASDAQ Exchange Timestamp.



Off-exchange trades are reported to the SIP quickly by the TRF. As Figure 4 depicts, the average time for a NASDAQ TRF report to reach the SIP is only slightly slower than the time it takes for a NASDAQ exchange report to reach the SIP. The time it takes for an off-exchange trade report to reach the TRF, however, is dramatically longer. Figure 3 reports this latency between the off-exchange trade execution timestamp and the TRF timestamp. There is considerable variation in the FINRA member to TRF, reflecting the diversity of FINRA members. The median latency is around 2,500 microseconds, which vastly dwarfs the sub-1,000 microsecond latencies between the TRF and SIP. The total time from off-exchange trade to SIP publication, therefore, is several thousands of microseconds larger than the on-exchange to SIP publication time. As Figure 11 documents, however, some novel information content of the off-exchange trade persists, with the spike in trades and quotes occurring in response to the off-exchange publication by SIP.

IV. Market Activity Analysis

A. Quote and Trade Volume Estimates

The SIP publication of TRF trades leads to a sudden, rapid increase in trading as well as quoting activity on the exchanges. The quantity and total volume of these responses is sizeable, as detailed in this section. Around 2% of dark trades in our sample have at least one on-exchange response trade, as detailed in Table I. These response trades occur with specific, persistent delays from the SIP publication time, with these delays lining up with geographic latencies of the exchanges.

Table I: Trading Response Probability. This table presents the probability that a given dark trade has a response trade on-exchange. That is, around 2% of dark trades published by the SIP have an apparent exchange response trade, which occurs at a delay which aligns with the relevant geographic latency. The relevant response regions for each exchange are highlighted in Figures 13 and 14.

Quote Volume

TRF	Tape A	Tape B	Tape C
NASDAQ TRF	1.8%	1.9%	2.2%
NYSE TRF	1.1%	1.1%	1.3%

The volume contained in the response trading bursts is sizeable. Table II provides detailed volume estimates for the volume of response trades. To calculate these volumes, we first calculate the total volume of on-exchange trading which occurs in the response bursts, as depicted in Figures 13 and 14. To eliminate the possibility that some of the apparent response trades occur by chance, we estimate a baseline level of trading volume by measuring total trading volume which occurs between 700 to 800 microseconds after the SIP publication of dark trades. We subtract this baseline level of volume from the total level of volume estimated in the response spikes and present the corrected figures in Table II. For the stocks in our sample, we observe approximately $775 billion in quoting volume and $65 billion in trading volume per year which occurs in the sharp response spikes to TRF trade publication. These volumes are the volume of the spike over and above the baseline estimate of volume.

Table II: Trading Response Volume. This table presents volume responses for our sample of stocks from January 1, 2019 to December 31, 2020. Responses are estimated for the specific latencies highlighted in Figures 13 and 14, and are the level of volume over and above the baseline region. All volumes are expressed in billions of dollars. These volumes over the two year sample amount to $1.5 trillion in quoting volume and $132 billion in trading volume, or $775 billion per year of quote update volume and $65 billion per year of trade volume.

(a) Panel A: Volume Responses to NASDAQ TRF Trades

Quote Volume

Exchange	Tape A	Tape B	Tape C
NASDAQ	209.1	89.8	147.0
CBOE	71.8	48.2	202.1
NYSE	102.6	100.0	249.2

Trade Volume

Exchange	Tape A	Tape B	Tape C
NASDAQ	8.5	5.9	33.2
CBOE	8.6	6.0	18.2
NYSE	8.1	4.7	21.5

(b) Panel B: Volume Responses to NYSE TRF Trades

Quote Volume

Exchange	Tape A	Tape B	Tape C
NASDAQ	53.4	13.7	49.2
CBOE	13.6	7.5	83.1
NYSE	19.2	12.9	80.9

Trade Volume

Exchange	Tape A	Tape B	Tape C
NASDAQ	0.7	0.6	6.4
CBOE	0.9	0.7	2.4
NYSE	1.0	0.6	4.1

B. Trade Analysis

While quotes are more numerous, trades show a commitment of capital to information. Trades also lead to real gains or losses for market participants involved in the transaction. By estimating realized spreads, we show that these trades are generally losses for the liquidity providers. In other words, traders executing such response orders to SIP publication of TRF trades gain profits, while the resting limit orders they execute against lose money. For each of the response ribbons identified in Figures 13 and 14, we estimate the mean realized spread for each stock-day observation in our

sample. We then average observations across each date, and report the estimates for each region in Table II. Realized spreads are consistently negative, suggesting that the response orders that execute earn money while the resting limit orders they execute against lose money (e.g., the response orders are able to "pick off" stale prices, as in Foucault, Roell and Sandas (2003)). Results are similar at the 30-second and 5-minute intervals, suggesting these gains are lasting trading profits that reflect underlying information content rather than transitory activity. These trading profits are earned on the response trades, which arguably reflects the informational value of the off-exchange trades, i.e., the information contained in these reports creates profitable trading opportunities for others. This highlights a potential mechanism how prices incorporate the information from retail trades that Barber, Lin and Odean (2021) identify. Boehmer, Jones, Zhang, and Zhang (2021) also offer suggestive evidence that retail orders contain information not incorporated into prices, and the reaction trades to TRF trade publications which we document provide further support for that hypothesis.

Table III: Trading Response Realized Spreads. This table presents mean realized spreads in basis points for trades in the response regions identified in Figures 13 and 14. Realized spreads compare the trade price with the mid-quote at some delay after the trade, and are a measure of the profit (for positive realized spreads) or loss (for negative realized spreads) available to a market maker. Spreads are averaged across stocks for each day, and then across days for each response region. All but two of the response regions of negative realized spreads, suggesting liquidity providers lose money on trades placed in these TRF-publication response trades. Results are consistent across both 30 second and 5 minute realized spreads. Response are in the same direction as the off-exchange trades, with buys following buys and sells following sells.

(a) Panel A: Realized Spreads (in Basis Points) for Trade Responses

Realized Spreads

Time Horizon	1st Quartile	Mean	3rd Quartile
30 Second	-0.53	-.33	-0.12
5 Minute	-0.52	-.20	0.13

(b) Panel B: Trade Direction for Trades and Quotes. For quotes, a buy sign would reflect reduced depth at the ask or a higher price at the ask, while a sell sign would reflect reduced depth at the bid or a lower price at the bid.

Percentage Trades with Same-Direction

	1st Quartile	Mean	3rd Quartile
Trades	64%	65%	67%
Quotes	59%	61%	64%

The profitability of these trades, in turn, suggests that the quotes are more than a mere passive incorporation of information. Rather, the SIP publication of TRF trades leads to profitable trading

opportunities at prevailing quotes, and a race between the traders who would snipe these stale quotes and those who would update their quotes before they are sniped. Participants on either side of this race would be willing to spend to obtain a latency advantage in this race, as it would allow them to earn profits or avoid losses.

C. Sub-penny Analysis

Off-exchange trades can be divided into three groups: midquote trades, sub-penny trades, and even penny trades. Boehmer, Jones, Zhang, and Zhang (2021) argue that sub-penny trades correspond to retail trades which receive *de minimis* price improvement to satisfy legal requirements. The majority of off-exchange trades are priced in even-penny increments, with over 70% of all off-exchange trades falling into this category. Midquote trades (classified as trades which have a sub-penny price between 40 and 60 hundredths of a penny) are the second-most common, comprising 19% of trades. Sub-penny trades (classified as trades with a sub-penny price between 1 and 39 or 61 and 99 hundredths of a penny) are the smallest category, comprising 11% of all trades.

For the even penny trades, subsequent on-exchange trades are even more strongly same-sign: 71% of the on-exchange response trades have the same trade direction as TRF trades. For midquote trades, the same-sign percentage is below 50%, but there is ambiguity in how midquote trades are signed. Since they do not happen at the bid or ask, it is not clear which party to the trade is taking or providing liquidity. Even-penny response trades earn largely negative realized spreads, while the midquote and sub-penny trades pay positive or zero realized spreads.

Table IV: Trade Responses and Off-Exchange Trade Type. This table analyzes how the lit response trades vary based on the type of off-exchange trade being reported. Off-exchange trades are divided into penny, sub-penny, and midquote trades. Penny refers to trades with a price at an even penny, midquote refers to trades with a sub-penny price between 40 and 60 hundredths of a cent, and sub-penny refers to the remainder of trades, i.e. those with a sub-penny price between 1 and 39 or 61 and 99 hundredths of a cent.

Response Trade Summary Statistics by Off-exchange Trade Type

Off-Exchange Type	Even Penny	Midquote	Sub-penny
Share of Trades	70%	19%	11%
Probability of Response	2%	2%	1.7%
Same-Share Percentage	71%	42%	59%
Mean Realized Spread (BP)	-0.57	0.3	0.05

D. Regression Analysis

We examine key drivers of the TRF-report response to volume spikes. Regression 1 estimates the relationship between TRF share and spike volume in the cross section of stocks. Spike volume is measured as the volume response on each exchange for the response ribbons identified in Figures 13 and 14, which capture the trading volume which lines up with the geographic latency for a SIP broadcast of an off-exchange trade report. We use TAQ data on the entire sample of Tape C Securities (i.e. NASDAQ listed) from January 1, 2019 to December 31, 2020. In this sample, the median stock has 0.20% of its daily volume occurring in this very narrow region which lines up with the latency-response to a TRF trade, where we measure this daily volume after controlling for the baseline level of expected trading. Given that this interval is less than 100 microseconds wide, 0.20% is a substantial share of the daily volume to occur in such a small interval.

Results of Regression 1 are presented in Table V. For the full sample of stocks, a higher TRF share in the cross section is associated with lower spike volumes, though this is driven by stocks which have very little exchange trading. If the sample is restricted only to securities which have at most 50% of their average daily volume reported through TRFs, a higher TRF share is associated with higher spike volume. Given the average response volume in the spike of 0.20%, an increase in TRF share from 30% to 35% is associated with a 3% increase in the daily volume that trades in the volume response spike.

REGRESSION 1: *For each stock **i**, we estimate:*

$$Volume\,Response\ Share_i = \alpha_0 + \alpha_1 TRF\,Share_i + \alpha_2 Price_i + \alpha_3 Vol_i + \alpha_4 \text{Spread}_i + \varepsilon_i \quad (1)$$

where Volume Response Share is the mean share of the stock which trades in the response spikes, TRF Share is the mean share of the stock reported through TRFs, Price is the mean nominal price, Volume is the mean nominal volume, and spread is the mean quoted spread.

REGRESSION 2: *For each stock **i** on date **t**, we estimate:*

$$Volume\,Response\ Share_{it} = \alpha_1 TRF\,Share_{it} + \alpha_2 Abs\ Overnight + \alpha_2 \text{Abs Intraday Return} + \alpha_3 \text{Spread} + \alpha_4 X + \varepsilon_{it} \quad (1)$$

where Volume Response Share is the mean share of the stock which trades in the response spikes, TRF Share is the share of the stock reported through TRFs, Volume is the nominal volume, Abs Overnight is the overnight return in absolute value, Abs Intraday Return is the intraday return in absolute value, and spread is alternately quoted, effective, or realized spread. X includes a fixed effect for each stock and each date, and standard errors are clustered by stock.

In the full sample of Tape C Securities, we have 4,129 stocks. Estimates from the full cross section give the reverse pattern: larger TRF share of trading volume is associated with *lower* trading volume in the response spike. This pattern is driven by securities where the majority of trades are reported through the TRF,[28] and for which exchange trading is less consequential. Due to the latencies involved with TRF reporting, and the lack of precision of the off-exchange timestamps, we cannot directly measure the off-exchange response to off-exchange trade reports.

We examine the pattern of volume response share and TRF share in the time series with Regression 2. To avoid the complication of very high TRF shares noted above, we restrict analysis to the sample of the 100 most traded Tape C securities. Results of Regression 2 are presented in Table VI. Results are similar to the cross section when it is restricted to the same subsample of the 100 most actively traded securities, with higher TRF share being associated with a larger volume spike response, and larger spreads being associated with a smaller volume spike response. Thus for securities with a TRF share below 50%, higher TRF share is associated with a larger spike response in both the cross section and the time series, suggesting that as the volume of TRF reports

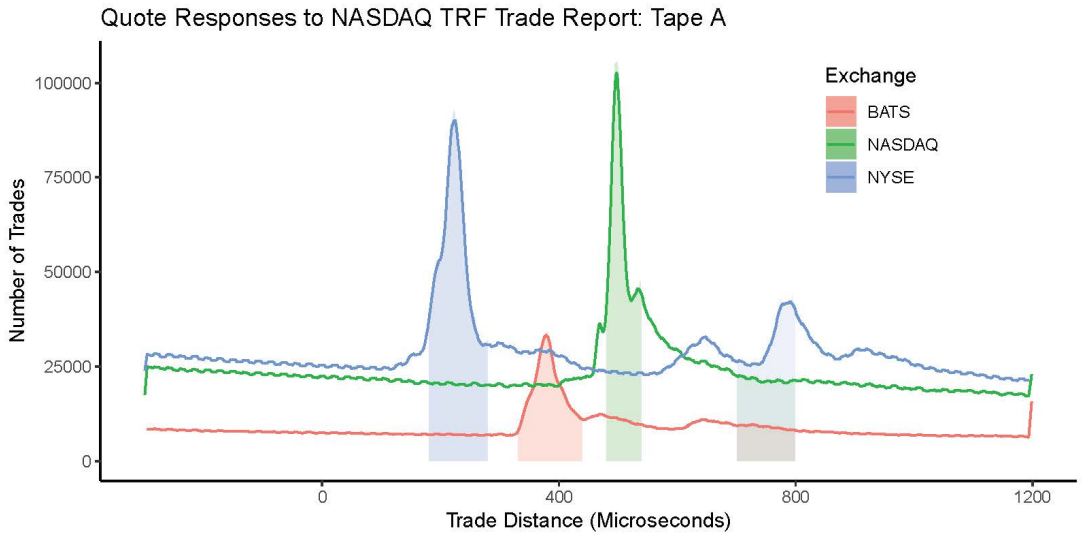
[28] In the sample of Tape C Securities, 306 securities have a TRF share between 75 and 100% while another 779 securities have a TRF share between 50 and 75%

grows, the trading response to the reports also grows. A 1% increase in off-exchange trading share leads to a 0.1% increase in on-exchange response volume.

Daily average spreads and response volume are inversely related. Wider quoted spreads, wider realized spreads, and wider effective spreads are all associated with a statistically significant decrease in the share of daily trading volume executed in the volume response spike. A one basis point increase in the quoted spread leads to a 0.05% decrease in spike volume share, while a one basis point increase in effective or realized spread leads to a 0.2% decrease in spike volume share. Li, Ye, and Zheng (2021) highlight the connection between low-latency races and spreads: they find that many races are driven by races around the minimum tick size. Our results identify a specific cause of some races: the race to respond to the publication of off-exchange trades. Consistent with their results, we find that the volumes in this race are tied to spreads, with larger spreads associated with fewer races.

Figure 13. Volume Estimates. Quote volume spikes on exchanges following the publication of a TRF trade report. Time zero denotes the SIP publication time of the TRF trade report. The x-axis measures the time between the SIP TRF timestamp, and the exchange timestamp of quotes. The y-axis denotes the total volume of quotes occurring at each possible offset. Ribbons highlight the area of response at each exchange. The ribbon from 700 to 800 microseconds highlights a baseline control region: this baseline estimates the total number of quotes that are expected to occur due to chance or bunching of quotes.

Panel A: Quote Responses to NASDAQ TRF Report: Tape A. Quote responses are estimated at 180-280 microseconds for NYSE, 330-430 microseconds for BATS, and 480 to 540 microseconds for NASDAQ. All baseline estimates are taken from 700 to 800 microseconds. Upgrades to the CTA SIP change these windows during our sample period. Figure 14 provides further details about the moving window which tracks these upgrades.



Panel B: Quote Responses to NASDAQ TRF Report: Tape C. Quote responses are estimated at 350 to 500 microseconds for NYSE, 180 to 260 microseconds for BATS, and 20 to 60 microseconds for NASDAQ. All baseline estimates are taken from 700 to 800 microseconds.

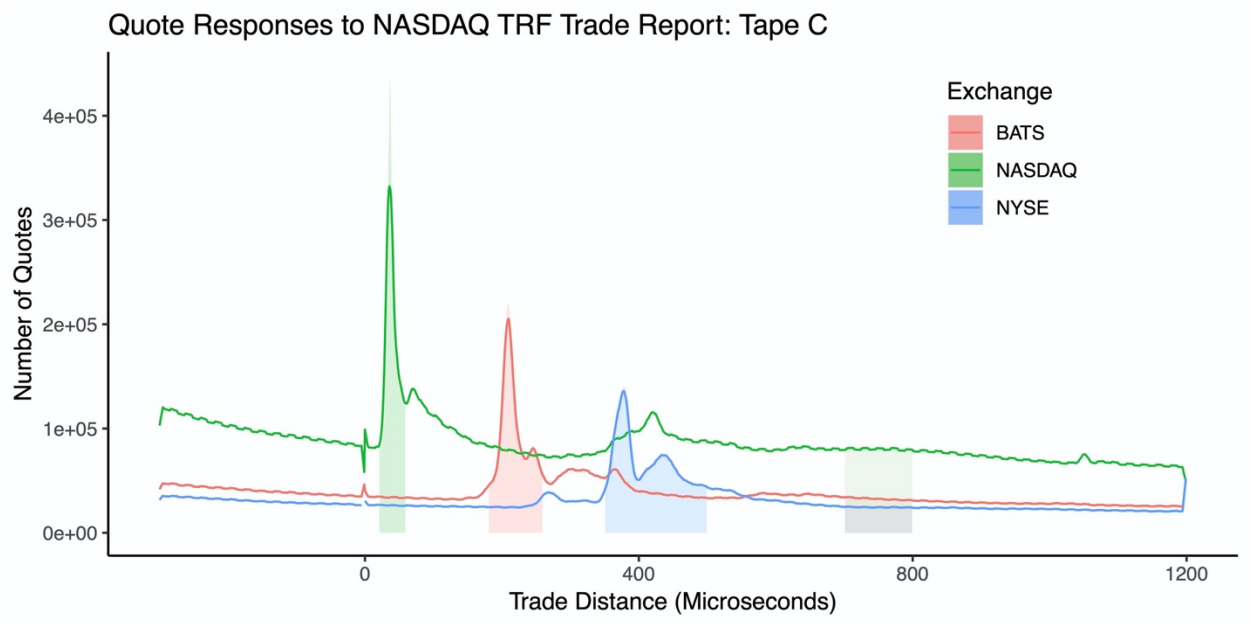


Figure 14. CTA SIP Moving Window. As depicted in Figure 4, Panel A, there have been several improvements to the CTA SIP over time. While the geographic location of the SIP has not changed, the order processing time has changed. Publication of new trades along with calculation of the NBBO has sped up considerably, and thus the time delay between the SIP timestamp and the arrival of the information at equity exchanges has decreased.

As an illustrative example, we present the quote-response pattern between the SIP timestamp and NYSE quoting volume for three distinct months in our sample. Across these months, time zero denotes the SIP publication time of the TRF trade report. The x-axis measures the time between the SIP TRF timestamp, and the exchange timestamp of quotes. The y-axis denotes the total volume of quotes occurring at each possible offset.

The trade response regions given in Figure 13 (a) are therefore modified to account for these changes in CTA SIP latency. We change the NYSE window to be 140 to 200 microseconds from September 2019 to May 2020, 120 to 180 microseconds from May 2020 to September 2020, and 20 to 50 microseconds from October to December 2020. The BATS interval falls to 240 to 320 microseconds, and the NASDAQ window falls to 370 to 450 microseconds.

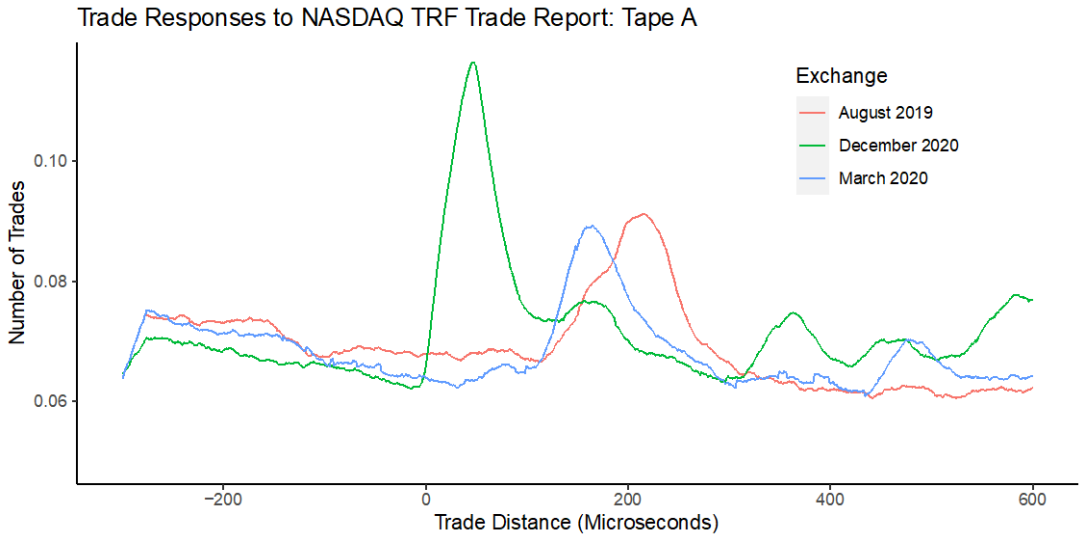


Table V: TRF Response Volume in the Cross Section. Regression 1 estimates relationship between Volume Response Share and TRF Volume Share, Price, and Spread. Observations are at the stock level. Volume Response Share is the mean share of the stock which trades in the response spikes in hundredths of a percent, TRF Share is the mean share of the stock reported through TRFs, Volume is the mean nominal volume, and spread is the mean quoted spread. The sample of stocks is all Tape C Securities from January 1, 2019 to December 31, 2020. Columns (1) and (2) estimate the regression only on stocks with a TRF Share of less than 50% of total trading volume, while columns (3) and (4) estimate the regression for the full sample of stocks.

	Dependent variable: Spike Share			
	TRF Share <50%		All Securities	
	(1)	(2)	(3)	(4)
TRF Share	0.057***	0.043***	−0.106***	−0.090***
	(0.019)	(0.016)	(0.008)	(0.007)
Price	0.022***	0.018***	0.018***	0.014***
	(0.003)	(0.002)	(0.003)	(0.002)
Quoted Spread		−0.022***		−0.017***
		(0.001)		(0.001)
Volume (Nominal)	−0.002***	−0.003***	−0.002***	−0.002***
	(0.001)	(0.0004)	(0.001)	(0.0004)
Constant	14.757***	17.591***	19.179***	20.888***
	(0.591)	(0.542)	(0.377)	(0.351)
Observations	2,126	2,071	2,952	2,852
R^2	0.033	0.159	0.080	0.168
Adjusted R^2	0.032	0.158	0.079	0.167
Residual Std. Error	9.906 (df = 2122)	7.957 (df = 2066)	9.325 (df = 2948)	7.893 (df = 2847)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table VI: TRF Share and Spike Volume. Regression 2 estimates relationship between Volume Response Share and TRF Volume Share, Returns, and Spreads. Observations are at the stock-day level. Volume Response Share is the share of the stock which trades in the response spikes in hundredths of a percent, TRF Share is the share of the stock reported through TRFs, Volume is the daily nominal volume, Abs Overnight is the overnight return in absolute value, Abs Intraday Return is the intraday return in absolute value, and spread is daily mean quoted, effective, or realized spread. The sample of stocks is the 100 most actively traded Tape C Securities from January 1, 2019 to December 31, 2020. We estimate a fixed effect for each stock and date and cluster standard errors by stock.

	Dependent variable: Spike Share		
	(1)	(2)	(3)
TRF Share	0.111^{***}	0.113^{***}	0.117^{***}
	(0.024)	(0.024)	(0.024)
Volume (Nominal)	0.0001	0.0001	0.0001
	(0.0001)	(0.0001)	(0.0001)
Absolute Overnight Return	11.549^{**}	11.778^{**}	10.934^{**}
	(4.725)	(4.689)	(4.808)
Abs Intraday Return	-0.997	-1.488	-1.220
	(2.752)	(2.755)	(2.755)
Quoted Spread	-0.057^{**}		
	(0.023)		
Effective Spread		-0.235^{***}	
		(0.065)	
Realized Spread (5s)			-0.267^{***}
			(0.060)
Observations	43,671	43,671	43,671
R^2	0.331	0.332	0.332
Adjusted R^2	0.322	0.322	0.322
Residual Std. Error (df = 43078)	11.804	11.799	11.799
Note:			*p<0.1; **p<0.05; ***p<0.01

V. Pre-Publication Trade Patterns

While the market is informed of off-exchange trades with a considerable delay, the participants in the off-exchange trade themselves will have knowledge of the trade before the broader market. This asymmetry in information does not exist for exchanges: when a trade is made on an exchange, the parties to the trade are notified no sooner than the public feed publication of the trade. For an off-exchange trade, the parties to the trade are potentially aware of the trade thousands of microseconds before the trade is published. We show that there is a considerable elevation of exchange trading around the time these off-exchange trades take place, indicating that at least one party to an off-exchange trade also trades on-exchange.

A. Simultaneous or Sequential Trading

When investors trade, they often make use of off-exchange and on-exchange facilities either uniquely or in combination. The order routing decision can be made by the investor directly, the investor indirectly through its use of a smart order router or be deferred to the broker. When an order-routing strategy relies upon the use both facilities, they can be used either *simultaneously* or *sequentially*. To access markets simultaneously, an order to buy or sell a security would be submitted to multiple facilities at the same time. With sequential access, an order is would be sent to one type of market facility and there it could execute completely or partially, or it could fail to execute. The party responsible for order routing would, at its discretion, send the remaining order flow to the other type of facility.

Sequential access can, potentially, make use of latencies in post-trade transparency. Suppose, for example, that an investor receives notification of a successful off-exchange trade. An investor can use this information to place an on-exchange trade, provided they are able to access the market between the time it takes for the trade report to transit the two-stage journey described in

Section II, with the trade reporting having to travel from off-exchange facility to TRF, and from TRF to SIP. As we document, after off-exchange trades are published there can be substantial revisions in quotes and trades. Trading before this trade publication allows an investor to benefit from the bid-ask spreads prevailing before their trade is published.

Accurate and consistent timestamps present a key challenge to identifying any trading activity that takes place across both lit and dark markets. Whether the strategy is simultaneous or sequential, the off-exchange trades will not face the same accuracy requirements in timestamp reporting. On-exchange trades have a regulatory requirement for microsecond or finer timestamps, and exchange clocks are required to be synchronized within 100 microseconds of the time maintained by the National Institute of Standards (NIST).[29] Off-exchange trades, in contrast, are only timestamped by the off-exchange facility to the nearest millisecond for TRF reporting.[30] This difference in granularity of the time stamp is important, as a single millisecond is sufficient time for a trade signal to transit the entire state of New Jersey. This causes a challenge when attempting to integrate trade and order data from the exchanges with the trade data originating from the TRFs. In this analysis, we continue to utilize the SIP timestamp of off-exchange trades, and examine trading activity (on and off exchange?) for up to 15,000 microseconds before the SIP timestamp of off-exchange trades. As Figure 3 highlights, the vast majority of (other?) off-exchange trades will occur less than 15,000 microseconds before the SIP timestamp.

[29] Cat NMS requires Participants to synchronize their clocks within 100 microseconds of the NIST time, with the exception of manual entry order events. NYSE TAQ trades are timestamped to the microsecond for NYSE and CBOE exchange trades, and to the nanosecond for NASDAQ exchange trades.

[30] FINRA Rule 6860 requires millisecond timestamps for Industry Members. Under FINRA Rule 4590, members are required to synchronize their clocks to a 50-millisecond tolerance of NIST time.

There is substantial trading reporting around the time off-exchange trades occur. Figure 14 shows the considerable trading activity which occurs between 0 and 15,000 microseconds prior to SIP publication of trades. As each off-exchange facility has a unique latency to the SIP, and the individual participant timestamps have only millisecond precision, we separately analyze the trading pattern around TRF trades according to their reported trade latencies. This method has as an underpinning the assumption that variation in trade reporting latencies will vary more across different reporters based on their geographic location, than within the trade reports by individual reporting brokers.

For each TRF trade, we measure the time difference between the participant and SIP timestamps. We divide TRF trades into latency categories depending on whether the latency is less than 1,000 microseconds, between 1,000 and 2,000 microseconds, between 2,000 and 4,000 microseconds, or greater than 4,000 microseconds. Figure 14 plots the pattern of exchange trading activity around the timing of trade publication for each latency category.

For trades reported between 1,000 and 2,000 microseconds (the green line of Figure 14), there are two clear spikes in trade activity. One spike occurs at approximately 800 microseconds prior to the SIP timestamp, i.e. at least 200 microseconds subsequent to the participant timestamp on the off-exchange trade. This is suggestive evidence for the subsequent-trading strategy, with orders executing off-exchange and then subsequently routing on-exchange before their off-exchange trade is published. One possible alternative explanation is that the timestamps are rounded up to the nearest millisecond, making the latency appear longer than it truly is.

B. Retail and Mid-quote Pre-Publication Trade and Quote Patterns

To further investigate how investors trade both on and off-exchange, we break trades out based on at the quote, mid-quote, and sub-penny price improvement of Boehmer, Jones, Zhang,

and Zhang (2021). In their framework, at the quote and midquote executions are associated with institutional order flow, while sub-penny price improvements are associated with retail order flow. Figure 15 depicts the trade and quote patterns around the different types of trade reports. For the midquote trades, there are substantially elevated trades and quotes, consistent with investors who are active in both markets. For the sub-penny trades, there is a substantial elevation in activity, but it is asymmetric between trades and quotes. For trades, there is substantial trading before the SIP publication, and a small sharp spike after trade publication. For quotes, the post-SIP publication quote updates are far more numerous, with a very large and sharp spike after the SIP publication. Prior to SIP publication, quoting volumes are elevated but not as comparatively high.

Investors who trade off-exchange may seek to change their on-exchange behavior due to both information and inventory considerations. For example, a wholesaler who internalizes an order to buy a security has both information from the customer's intent, and a long position in the security. Both active trades and passive quotes could both used to control inventory and execute profitable trades based on private information. Submitting a marketable order requires paying a spread, however, while a passive limit order gains the spread. If a wholesaler internalizes a trade, the price paid must be at least the NBBO, and in Figure 15, the trades are at a sub-penny improvement to the NBBO. To profitably unwind this position, then, a wholesaler must be able to trade at or better than the NBBO, which would require a passive quote. We instead observe a substantial elevation in trades, and not quotes, prior to the SIP publication of the trade, suggesting an informational motivation, and not an inventory motivation, to these trades.

Figure 14. Pre-Publication Trade Patterns By Latency. Prior to the publication of an off-exchange trade report, trading activity on exchanges is substantially higher, indicating investors are active in both exchanges and off-exchange facilities. Time zero denotes the SIP publication time of the TRF trade report. The x-axis measures the time between the SIP TRF timestamp, and the exchange timestamp of trades. The y-axis denotes the total volume of trades occurring at each possible offset. Each line depicts exchange trading relative to a specific subset of off-exchange trades, where each subset of off-exchange trades has a reported participant-to-TRF latency. The green line, for example, depicts the pattern of NASDAQ trading activity around the time that TRF trades are reported with a participant-to-SIP timestamp difference of 1,000 to 2,000 microseconds.

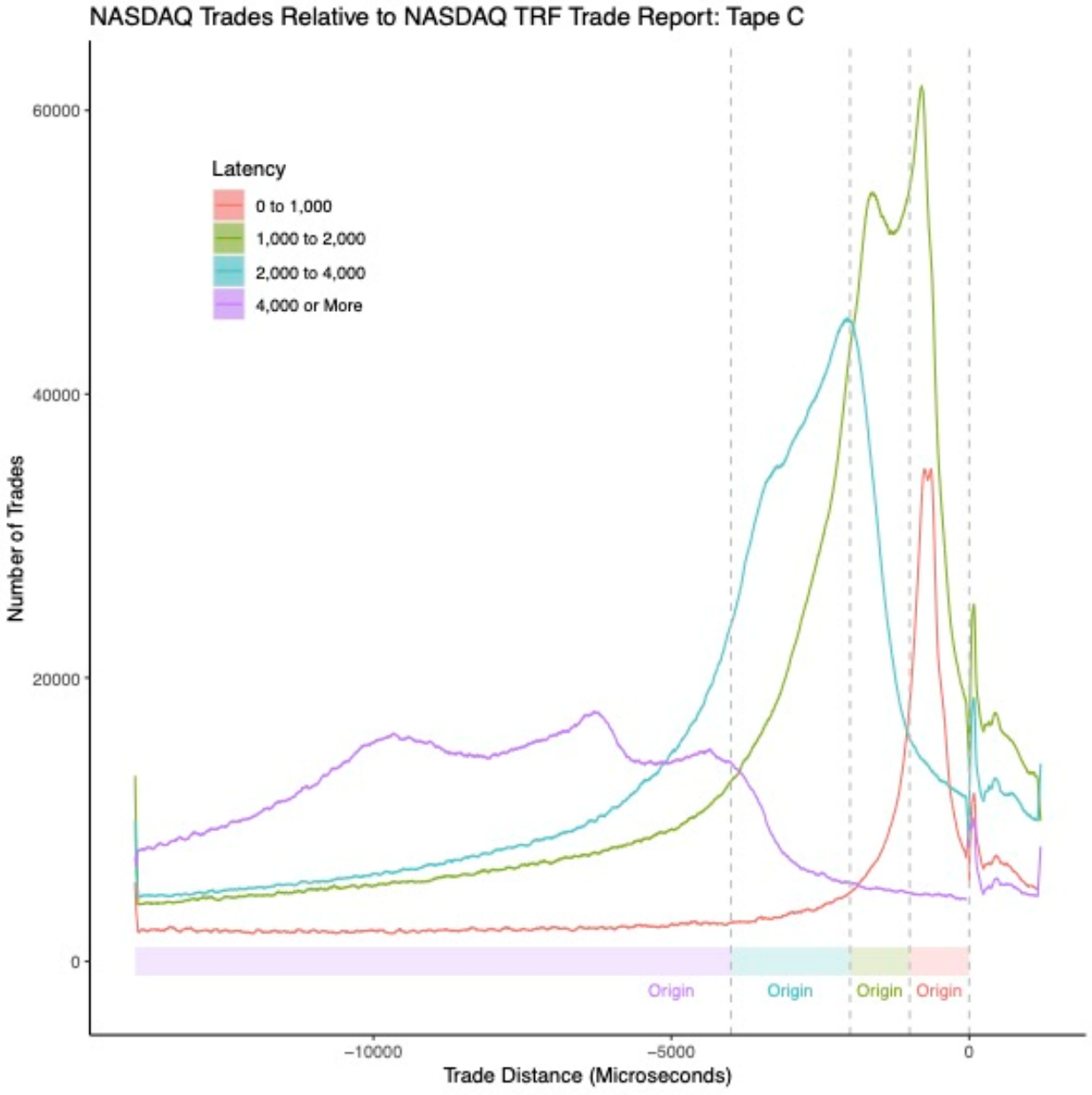
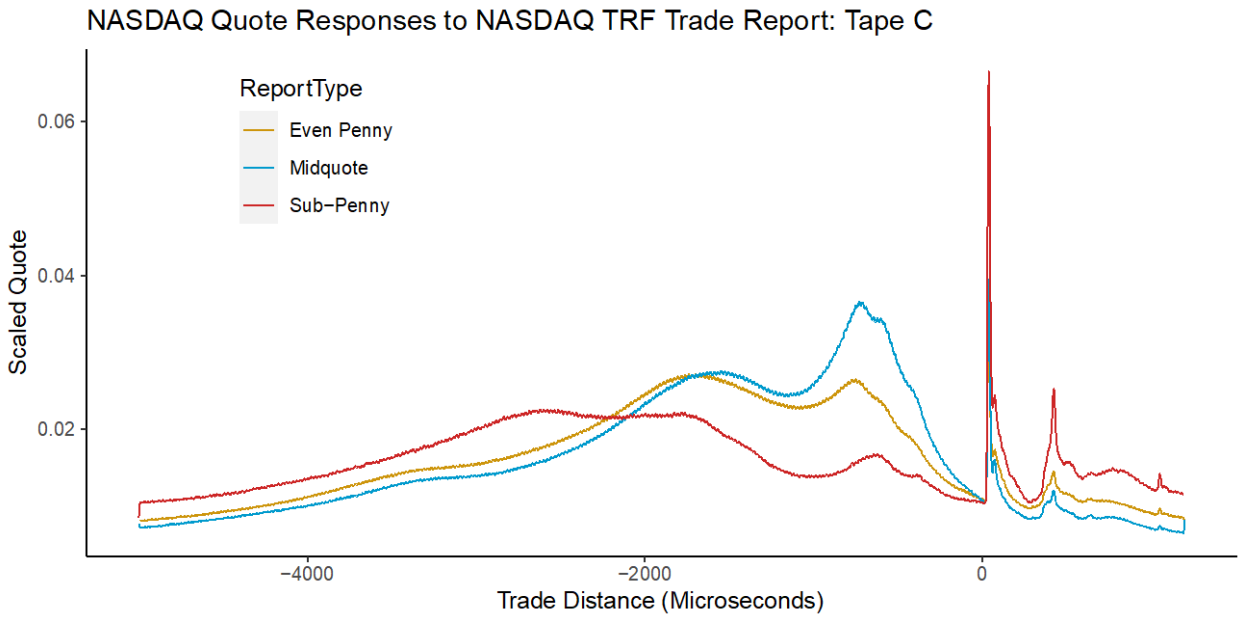
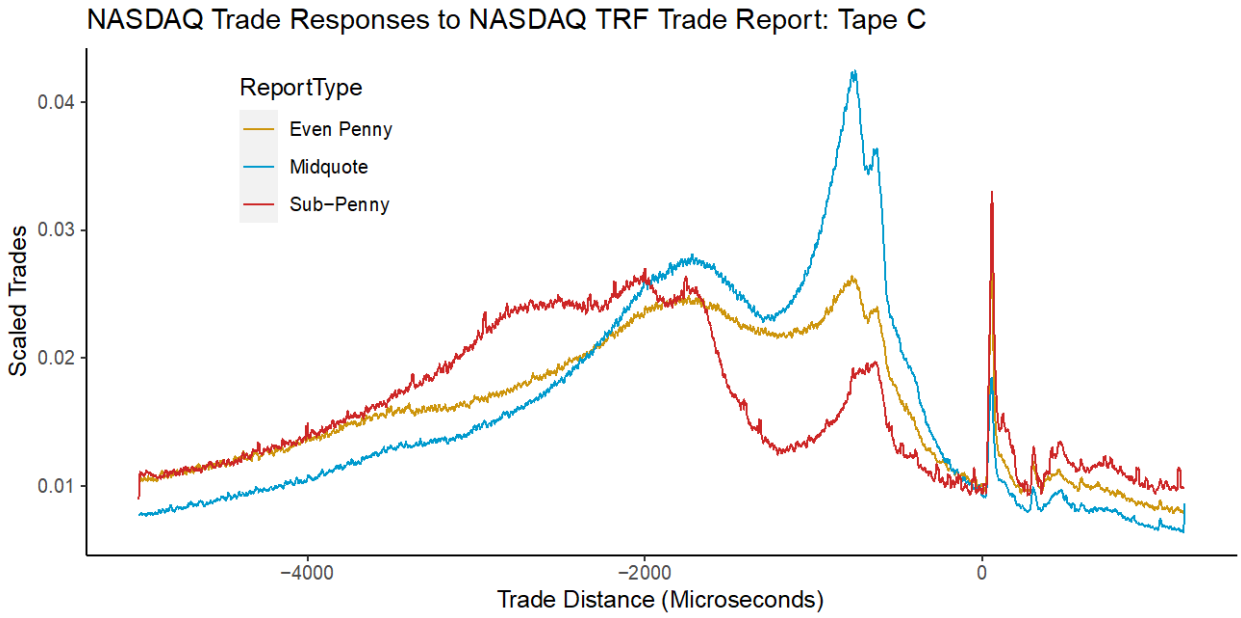


Figure 15. Pre-Publication Trade and Quote Patterns Around Trades. Quote volume spikes on exchanges around the publication of a TRF trade report, where the TRF trade report is priced at either the even penny, the midquote, or a sub-penny increment. Time zero denotes the SIP publication time of the TRF trade report. The x-axis measures the time between the SIP TRF timestamp, and the exchange timestamp of quotes. The y-axis denotes the total volume of quotes (Panel A) or trades (Panel B) occurring at each possible offset.

Panel A: Quote Pattern Around NASDAQ TRF Report: Tape C.



Panel B: Trade Pattern to NASDAQ TRF Report: Tape C.



VI. Conclusion

Off-exchange trading has become an increasingly important part of the equity market ecosystem. While publication of off-exchange trading by the SIP leads to a dramatic increase in trading and quoting in that market participants learn from and respond to these reports for off-exchange trades, there isn't an apparent response to exchange trading after publication by the SIP as the reaction occurred earlier due to the proprietary feeds. It is surprising and puzzling that even though the off-exchange trades are included in particular proprietary feeds (though not the expensive feeds oriented to high frequency trading firms), market participants do not react to the publication of off-exchange trades until the publication by the SIP.

Exchange and off-exchange trading have a number of different features. For example, the off-exchange trading is relatively more attractive to retail investors, a category of investors attracting attention in the aftermath of the pandemic (in general) and GameStop and other meme stocks (in particular). Also, off-exchange data is focused on transaction prices, while the proprietary data that emanates from exchange markets would include quotes and orders. Of course, there are interconnections at many levels between exchange and off-exchange trading (beyond the direct trading interactions and regulatory obligations, such as "Best Execution" and Regulation NMS). For example, the Trade Reporting Facilities for off-exchange trading are operated by two of the three major exchange affiliate families (NYSE and NASDAQ) and important market makers operate in both exchange and non-exchange venues (per differing regulatory treatments). Our paper highlights the faster effective responses and reporting to exchanges rather than off-exchange data. Traditionally, the focus on opacity of off-exchange trading has focused on pre-trade opacity, but our analysis highlights that post-trading reporting is slower in off-exchange contexts (relative to post-trade opacity of off-exchange trading, despite the presence of a reporting obligation in equity markets).

Our analysis also provides an interesting approach to highlight the potential value of data by identifying the increase in activity that the availability of particular data induces and the endogenous price advantage that is reflected in the resulting size of the negative effective spread.

Methodologically, we point to the usefulness of the data and informational flows from various venues and tie this to the geographical structure of latency using publicly available TAQ data.

In recent years there has been considerable regulatory attention as to the appropriate pricing of market data as the pricing of exchange proprietary data has been quite contentious. In fact, the SEC has moved forward to adopt changes to the regulatory framework for consolidated (SIP-like) data. We document that while there are no market reactions to SIP publication of on-exchange trades (due to the existence of faster proprietary feeds), there is a sharp, sudden market response to the SIP publication of off-exchange trades. These response trades appear informed: they are in the same direction as the off-exchange trades, and earn negative realized spreads on average. Whether the off-exchange price is in an even penny (suggestive of institutional trades) or a sub-penny increment (suggestive of a retail trade), we document a similar pattern.

REFERENCES

Aquilina, Matteo; Eric Budish, and Peter O'Neill, 2020, "Quantifying the High-Frequency Trading 'Arms Race': A Simple New Methodology and Estimates," Forthcoming in *Quarterly Journal of Economics.*

Barber, Brad M., Shengle Lin, and Terrance Odean, 2021, "Resolving a Paradox: Retail Trades Positively Predict Returns but are Not Profitable," working paper.

Bartlett, Robert P., and Justin McCrary, 2019, "How Rigged are Stock Markets? Evidence from Microsecond Timestamps," *Journal of Financial Markets* 45, 37-60.

Boehmer, Ekkehart, Charles M. Jones, Xiaoyan Zhang, and Xinran Zhang, 2021, "Tracking Retail Investor Activity," *Journal of Finance*, forthcoming.

Budish, Eric; Peter Cramton, and John Shim, 2015, "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Quarterly Journal of Economics* 130, 1547-1621.

Comerton-Forde, Carole, and Talis J. Putnins, 2015, "Dark Trading and Price Discovery," *Journal of Financial Economics* 118, 70-92.

Conrad, Jennifer, & Sunil Wahal, 2020, "The Term Structure of Liquidity Provision," *Journal of Financial Economics* 136, 239-259.

Cox, Justin, 2019, "NASDAQ and the NYSE: A Trade Reporting Facility Comparison," Appalachian State University.

Foucault, Thierry, Ailsa Roell, and Patrik Sandas, 2003, "Market Making with Costly Monitoring: An Analysis of the SOES Controversy," *Review of Financial Studies* 16, 345-384.

Hasbrouck, Joel, 2019, "Price Discovery in High Resolution," *Journal of Financial Econometrics*, 1-36. https://doi.org/10.1093/jjfinec/nbz027.

Holden, Craig W., and Stacey Jacobsen, 2014, "Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions," *Journal of Finance* 69, 1747–1785.

Hu, Grace Xing, Jun Pan, and Jiang Wang, 2017, "Early Peek Advantage? Efficient Price Discovery with Tiered Information Disclosure," *Journal of Financial Economics* 126, 399-421.

Kim, S., & Trepanier, D. (2019). Violations of Price-Time Priority and Implications for Market Quality. Available at SSRN 3507106.

Li, Sida; Mao Ye; and Miles Zheng. "Refusing the Best Price". Working Paper, University of Illinois, Urbana-Champaign.

Seppi, Duane, 1990, "Equilibrium Block Trading and Asymmetric Information," *Journal of Finance* 45, 73-94.

Spatt, Chester S., 2021, "Is Equity Exchange Market Structure Anti-Competitive"? Working Paper, Carnegie Mellon University.

Zhu, Haoxiang, 2014, "Do Dark Pools Harm Price Discovery?" *Review of Financial Studies* 27, 747-789.