How Rigged Are Stock Markets?
Evidence from Microsecond Timestamps

Robert P. Bartlett, III*
University of California, Berkeley

Justin McCrary**
University of California, Berkeley, NBER

Abstract:

Using new data from the two Securities Information Processors (SIPs), we examine claims that HFT firms use direct feeds to exploit traders who rely on SIP prices. Across $3 trillion of trades, the SIPs report quote updates from exchanges 1,130 microseconds after they occur. However, the SIP-reported NBBO matches the NBBO calculated without reporting latencies in 97% of all SIP-priced trades. Liquidity-taking orders gain on average $0.0002/share when priced at the SIP-reported NBBO rather than the instantaneous NBBO, but aggregate gross profits are just $11.6 million. These findings indicate that direct feed arbitrage is not a meaningful source of HFT profits.

Draft Date: July 28, 2017
JEL codes: G10, G15, G18, G23, G28, K22
Keywords: latency arbitrage, high-frequency trading; SIP; market structure

***Statement of Financial Disclosure and Conflict of Interest*: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.**

* rbartlett@berkeley.edu, 890 Simon Hall, UC Berkeley, Berkeley CA 94720. Tel: 510-542-6646. Corresponding author.
** jmccrary@berkeley.edu, 586 Simon Hall, UC Berkeley, Berkeley CA 94720. Tel: 510-643-6252.

"Some have suggested that exchanges that use the SIP data to calculate the NBBO provide unfair opportunities to sophisticated traders to engage in risk-free latency arbitrage."

- *Senate Testimony of Joseph Ratterman, Chief Executive Officer of BATS Global Markets, June 14, 20014*

1. Introduction

Concerns over the different speeds at which market participants access information and the resulting potential for adverse selection in equity markets have occupied center stage in recent years. In particular, the emergence of low-latency trading strategies that can exploit sub-second information asymmetries has led not just to economic research, but also to extensive regulatory scrutiny, litigation, and the approval in 2016 of the Investors Exchange (IEX) as a new stock exchange. Describing high frequency trading (HFT) as "one of the greatest threats to public confidence in the markets," New York attorney general Eric Schneiderman in 2014 launched a series of high profile lawsuits against dark pools, exchanges, and HFT firms. Regulators from the Federal Bureau of Investigation,[1] to the Commodity Futures Trading Commission,[2] to the Securities and Exchange Commission (SEC) have all brought pressure to bear on HFT.[3]

Within this debate, an especially important flashpoint has emerged regarding the differing speeds at which traders can access and process data emanating from approximately one dozen U.S. stock exchanges. For instance, the controversial use by IEX (and planned use by the NYSE MKT and the Chicago Stock Exchange) of so-called "speed bumps"— intentional delays between the time an order is entered on an exchange and the time it is executed or posted—is rooted in a desire to level the playing field between fast traders having preferential access to exchanges' quotation data and other traders on the venue. Similar concerns about fast traders' preferential access to exchange quotation data motivated the

[1] Scott Patterson and Michael Rothfeld, "FBI Investigates High-Speed Trading," *Wall Street Journal*, March 31, 2014. Available at http://www.wsj.com/articles/SB10001424052702304886904579473874181722310, last accessed December 17, 2016.
[2] Douwe Miedema, "U.S. Futures Regulator CFTC Probing Speed Traders," *Reuters Business News*, April 3, 2014. Available at http://www.reuters.com/article/us-hedgefunds-speed-trading-cftc-idUSBREA321QU20140403, last accessed December 17, 2016.
[3] John McCrank, "Exclusive: SEC Targets 10 Firms in High Frequency Trading Probe—SEC Document," Reuters Business News, July 17, 2014. Available at http://www.reuters.com/article/us-sec-investigation-highfrequencytradin-idUSKBN0FM2TW20140717, last accessed December 17, 2016.

SEC's widely-followed investigation in 2016 of the market-making firm Citadel Securities (Levinson, 2016).

In general, these concerns arise from the institutional fact that trading rules generally require brokers and trading venues to fill market orders at (or better than) the national best bid or offer (the "NBBO") available across exchanges. Indeed, many venues—particularly non-exchange venues—expressly price transactions by "pegging" them to the NBBO. Market participants can determine the NBBO by looking to its publication by the two centralized Securities Information Processors ("SIPs") to which all exchanges are required to report updates to their best bids and offers; however, exchanges are also permitted to provide their quote updates directly to subscribers using superior data feeds. If exchanges provide fast traders with the ability to calculate the NBBO microseconds before other traders relying on the SIPs or other slower data feeds, exchanges are effectively allowing fast traders to foresee changes to the NBBO on which other traders will be transacting, potentially enhancing these traders' adverse selection costs. Because these fast traders would exploit the speed advantage of buying the fastest quote data from exchanges rather than relying on slower data feeds from the SIPs, we refer to this trading behavior as "direct feed arbitrage."

Until recently, understanding the extent to which traders engage in direct feed arbitrage has been hampered by the absence of detailed information concerning the informational advantage of fast traders who obtain exchange data from exchanges' proprietary data feeds rather than the SIPs. In the meantime, concerns that a principal source of HFT rents comes from exploiting these informational advantages has shaped the broader debate concerning the welfare consequences of the arms race for trading speed. By paying for faster access to exchange trading data, do HFT firms obtain rents in the form of risk-free arbitrage? Or are these concerns just a distraction from understanding the primary sources of HFT profits, whether benign, such as conventional market-making (Menkveld, 2013), or not (such as quote-stuffing (see, for example, Egginton, Van Ness, and Van Ness, 2016))?

In this paper, we use new timestamp data provided by the two SIPs to conduct the first market-wide analysis of the latency with which the SIPs process quote and trade data, and we present new results regarding the economic significance of direct feed arbitrage. These data are the result of a regulatory change obligating exchanges and broker-dealers to report to the appropriate SIP the precise time (measured in microseconds) at which a trading venue either updated a quotation or executed a trade. Moreover, amendments to the SIP operating procedures at this time required the two SIPs to record in microseconds the precise time at which each SIP processed a trade or quotation update submitted by an exchange or broker-dealer. Comparing these two timestamps thus permits a direct analysis of the SIP processing latency for all trades and quote updates across the entire market. For ease of computation, we focus on all trades involving the Dow Jones 30 during the first eleven months of these new reporting requirements.[4]

To preview our specific findings, we first document descriptively that the mean time gap between the time a quote update is recorded by an exchange matching-engine and the time it is processed by a SIP is now just 1,130 microseconds. Mean latency for processing trades, however, is approximately 20 times higher, clocking in at 22,840 microseconds. Due to these reporting latencies, we show that the NBBO reported by the SIP lagged the "true" NBBO on average 6,839 times per day for the Dow Jones 30.

In addition to describing these new data, we use them to explore empirically the economic significance of direct feed arbitrage. We focus on the costs of trading at stale SIP prices for liquidity takers and for liquidity providers. Somewhat surprisingly, both classes of traders are commonly alleged to be injured by direct feed arbitrage, often at the hand of the other. For instance, the SEC's 2016 investigation into the retail market-making firm Citadel Securities focused on the allegation that market makers filling marketable orders at (or within) the SIP-generated NBBO often do so at stale prices to the disadvantage of retail investors using marketable orders.[5] At the same time, the premise behind the "speed bumps" at IEX and other exchanges is that liquidity providers need protection from the strategic

[4] As noted below, we also report extensions of selected key findings to half and three-quarters of the full equities market (by trading volume). These results are qualitatively similar to our results using the Dow Jones 30 alone.
[5] For instance, suppose a direct feed showed the NBBO changing from $10.00 x $10.01 to $9.99 x $10.00, while the SIP's NBBO remained at $10.00 x $10.01. A broker might fill buy orders by selling to them at $10.01 (the stale NBO reflected in the SIP NBBO) rather than at $10.00 (the NBO shown in its direct feed). We discuss the Citadel case in more detail in Section 5(a).

use of marketable orders by HFT firms to "pick off" resting limit orders that have been pegged to stale NBBO prices.[6]

While the first strategy has not been studied in the academic literature, the latter strategy is consistent with prevailing models of HFT which examine how the presence of fast traders can raise adverse selection costs for dealers and slower traders using limit orders.[7] At the same time, however, Hoffmann (2014) demonstrates that the very risk of being adversely selected produces strong incentives for liquidity providers to invest in speed to avoid quoting at stale prices. Brogaard, et al. (2015) show empirically that market-makers are especially inclined to invest in faster technology. Together, these papers suggest liquidity providers will trade at venues that rapidly update their estimation of the NBBO, which will limit the opportunity to trade against liquidity providers at stale SIP prices.

To estimate empirically how much traders lose by trading at stale SIP prices, we examine how liquidity takers and liquidity providers fared by trading at prices matching the SIP-generated NBBO rather than the NBBO calculated in a world without any reporting latencies. In general, we ask the following hypothetical: *If every trade occurring at a price equal to the SIP-generated NBBO reflects a trader being subject to adverse selection because of direct feed arbitrage, what are the maximum trading losses to liquidity takers? And what are the maximum trading losses to liquidity providers?* To answer these questions, we start by showing how to use the new timestamps reported to the SIPs to reconstruct for each trade in our sample the NBBO that prevailed on the SIP (the "SIP NBBO") at the microsecond in which the trade occurred, along with the NBBO that was theoretically possible were there no latency at all

---

[6] As an illustration of this behavior, consider the following example given in Fox, Glosten & Rauterberg (2015). In it, an institutional investor posts to a dark venue a midpoint buy order for a security when the NBBO is $161.11 x $161.15 so that an incoming market order to sell would result in this order being filled at $161.13. However, if the exchange holding the best ask subsequently decreases its displayed quote from $161.15 to $161.12 while the midpoint order rests in the dark pool, a fast trader can detect the new NBBO before the dark venue, providing it a momentary opportunity to send an immediate-or-cancel sell order to the dark venue that will execute at the stale midpoint of $161.13. Upon receiving confirmation, the fast trader can cover the resulting short position by sending a marketable buy order to an exchange to execute at the new national best bid of $161.12, producing a penny of risk-free profit. In the meantime, the institutional investor—rather than buying at $161.115, the actual midpoint—buys at $161.13.

[7] See, for instance, Brogaard et al. (2015) Budish, Cramton, and Shim (2015), Foucault, Hombert, and Rosu (2016), Hendershott and Moulton (2011), Hoffmann (2014), Jovanovic and Menkveld (2012), or Menkveld and Zoican (forthcoming).

in transmitting quote updates (the "Direct NBBO"). Reconstruction of this "direct feed" NBBO is made possible by the fact that for each quote update from an exchange, the new timestamp data includes the time at which a quote update was released by the exchange matching engine and therefore available for distribution over an exchange's direct proprietary data feed.

With these measures, we estimate over our sample period the gross profits gained and lost on each trade that matched the SIP NBBO rather than the Direct NBBO. Importantly, a trade price that matches a stale SIP price can arise either because a trading venue used the SIP NBBO to price a trade or because a trading venue used direct data feeds but was too slow to update its calculation of the new NBBO before the trade occurred. Regardless of why a trade price matches the stale SIP NBBO, the existence of these trades reveals an opportunity for a fast trader to profit from the ability to calculate the new NBBO faster than others in the market. Indeed, because the Direct NBBO assumes zero latency in transmitting and processing quote data as well as zero transaction costs, our methodology provides an outer maximum of the overall profitability to fast traders from trading with others at stale SIP prices. Moreover, because trade prices matching the SIP NBBO can arise from venues that actually rely on SIP data as well as venues that use direct data feeds (but process the data slowly), our approach permits insight into the profitability of direct feed arbitrage strategies despite the increasing use of direct data feeds by many venues.

Overall, our analysis suggests quote reporting latencies generate remarkably little scope for exploiting the informational asymmetries available to subscribers to exchanges' fastest direct data feeds, regardless of whether trading is targeted at liquidity takers or at liquidity providers. Indeed, with respect to liquidity takers, on a size-weighted basis, liquidity-taking trades in our sample that match either the SIP NBB or the SIP NBO would have *gained* on average $0.0002 per share by having their trades priced at the SIP NBBO rather than the Direct NBBO. This number is small in magnitude because approximately 97% of trades within our sample occur at a time when the SIP NBBO and Direct NBBO are the same pair of numbers. This simple fact highlights the low probability that the choice of NBBO benchmark matters at all for liquidity-taking trades at the best ask or best offer. Moreover, we show that when the SIP NBBO and Direct NBBO differ, liquidity taking traders systematically benefit by having their trades priced at the SIP NBBO. That liquidity takers gain on average is surprising in light of contemporary debates about equity market structure, but the finding makes sense: The NBBO will often change in response to serial buy (sell) orders so that late-arriving buy (sell) orders benefit from the stale quotes that have yet to reflect the new trading interest.

Because there are two sides to every trade, our finding regarding the benefits to liquidity takers of trading at SIP prices naturally raises the possibility that liquidity providers who trade at stale SIP prices are being "picked off" by fast traders to earn risk-free profits. To examine empirically whether this is the case, we exploit the fact that such an arbitrage play would require a pair of trades and would thus generate a data residue. We find little evidence that these trades are the result of fast traders using market orders to "pick off" stale limit orders priced at the SIP NBBO to earn risk-free profits. Specifically, our analysis shows that at most 0.8% of these liquidity-taking trades could be part of such a strategy.

Equally important, while our sample of SIP-priced trades amounts to nearly $3 trillion of transaction value, we estimate that a liquidity taker capable of picking off *every* stale quote at the SIP NBBO where doing so was advantageous to the liquidity taker would have earned just $11.6 million in gross profits before accounting for the costs of the second-leg transaction. By comparison, trading spreads for these stocks are usually near a penny, so the total trading spreads available to liquidity providers for these $3 trillion of trades were roughly $30 billion. Consequently, an HFT firm focused on simply earning the spread on these trades would be competing for gross profits that were well over 3,000 times as great as the gross profits available from an active strategy focused on picking off stale quotes at the SIP NBBO. This latter finding underscores how, at least in the present market, HFT strategies other than direct feed arbitrage are considerably more likely to account for the high speed arms race.

Nor does this conclusion change when we look beyond the Dow Jones 30 to estimate the aggregate profitability of these direct feed arbitrage strategies for the entire trading market. While the annual trading value of SIP-priced trades is over $30 trillion, our estimate of the maximum available profits liquidity providers could earn on these trades from direct feed arbitrage is less than $5 million per year. We similarly estimate the maximum amount of annual gross profits available from picking off stale quotes priced at the SIP NBBO is approximately $70 million before accounting for any second-leg trades or other trading costs.[8]

This paper is most closely related to two recent studies of latency arbitrage. Wah and Wellman (2013) estimate the prevalence of latency arbitrage opportunities created by market fragmentation when two or more exchanges create a crossed market (i.e., when the best bid on one exchange creates a NBB that is greater than the NBO). However, their analysis focuses on latency arbitrage strategies designed to exploit crossed markets, while we focus on strategies designed to exploit quote reporting latencies. More relevant to our empirical analysis of direct feed arbitrage is Ding, Hanna & Hendershott (2014) who study the latency between NBBO updates provided by the publicly-available SIP and NBBO updates calculated using proprietary data feeds for a trader based at the BATS exchange in Secaucus, New Jersey. For such a trader, they find that price dislocations between the two observed NBBOs average 3.4 cents and last on average 1.5 milliseconds. Using a single trading day for Apple, Inc., they use these estimates to conclude that a fast trader could theoretically earn up to $32,000 over the course of the trading day by trading against stale orders in dark pools based on the volume of off-exchange trades. This estimate, however, assumes each off-exchange trade is made during a period of price dislocation. Our data, in contrast, permits analysis of how many trades are actually made during a period of price dislocation across both exchange and non-exchange venues, enabling a precise estimate of the probability that a trade is adversely affected by direct feed arbitrage. Our data also permits an estimate of the trading gains and losses traders experience by having their trades priced at the SIP NBBO. Consequently, our results establish that such fast traders are not likely to be as highly compensated as the analysis in Ding, Hanna, and Hendershott (2014) suggests.

Finally, while our results establish that there is little scope in equity markets currently for direct feed arbitrage, we caution that these results should not be over-interpreted. In particular, our results *do not*

[8] By comparison, 2016 revenue for Virtu Financial, a single HFT firm subject to SEC reporting obligations, was nearly $700 million, suggesting the profitability of HFT is to be found outside these quote latency strategies.

rule out other types of latency arbitrage that might be prevalent in the current environment. Nor do our results rule out the possibility that direct feed arbitrage might have been prevalent in the quite recent past (e.g., 2014), for the simple reason that our data are not available until mid-2015. Nonetheless, our results do clarify that a popular narrative regarding direct feed arbitrage would appear to be scarcely relevant to markets in 2015-2016, and they provide the first broad-based evidence on the extent of quote, trade, and NBBO latency using the SIPs' new microsecond timestamps.[9]

The remainder of this paper is organized as follows. Section 2 provides institutional details regarding the rules governing the dissemination of trade and quote data and the advantage that direct feed data gives fast traders. Section 3 summarizes the new microsecond timestamps and sample selection choices. Section 4 presents our empirical estimates of trade and quote reporting latencies. Section 5 examines the economic consequence to liquidity takers and liquidity providers of having trades priced at the SIP NBBO rather than the Direct NBBO. Section 6 concludes.

2. Institutional Background

At present, there are three national market plans governing the dissemination of quote and trade data for National Market System (NMS) equity securities. These three plans are required by Rule 603 of Regulation National Market System (Reg. NMS) and reflect the historical structure of U.S. equity markets. For trades in NYSE-listed securities ("Tape A" securities) and securities listed on regional exchanges and their successors ("Tape B" securities), the Consolidated Trade Association ("CTA") Plan requires all exchanges and FINRA to report last sale information to the Securities Industry Automation Corporation ("SIAC"), a subsidiary of the NYSE which acts as the central SIP for any transaction in Tape A and Tape B securities. The Consolidated Quotation ("CQ") Plan similarly obligates exchanges and

[9] A related issue arising from the availability of direct data feeds concerns the fact that any divergence between the SIP NBBO and the NBBO derived from direct data feeds can create the possibility for conflicting trade execution measures depending on which NBBO a venue chooses to use as its pricing benchmark. For instance, in connection with government investigations into the use of direct feed data by Citadel's market-making division, Levinson (2016) reports concerns that Citadel's trade execution statistics are based on the slower SIP data rather than the NBBO available from direct data feeds. In an Internet Appendix, we provide an extensive analysis of how the choice of NBBO benchmark affects a venue's trade execution statistics. Calculating effective spreads using the Direct NBBO rather than the SIP NBBO changes effective spreads by less than 1.9 percentage points for exchange trades and less than a half percentage point for all non-exchange trades.

FINRA to report to the SIAC any change in the best bid or best offer (including changes to the number of shares) currently available on each trading venue for Tape A and Tape B securities, which the SIAC uses to calculate the NBBO for these securities.[10] For transactions in Nasdaq-listed securities ("Tape C" securities), the Unlisted Trading Privileges ("UTP") Plan governs reporting obligations for both trades and quotations. Under this plan, exchanges and FINRA must provide trade and quote updates in any Tape C securities to Nasdaq, which operates as the SIP for transactions in these securities. We refer to the SIP managed by the SAIC as the "NYSE SIP" and the SIP managed by Nasdaq as the "Nasdaq SIP."

While the trade reporting plans initially focused on exchange-based trades, the SEC has required since March 2007 that all off-exchange transactions be reported to a formal FINRA-managed Trade Reporting Facility (a "FINRA TRF") (O'Hara & Ye, 2011). At present, FINRA manages two facilitates operated separately by the NYSE and Nasdaq. In combination with FINRA's trade reporting obligations under the CTA and UTP Plans, this SEC reporting requirement for FINRA members means that off-exchange trades made through a broker-dealer internalizer or in a dark pool are now effectively segregated and reported to the appropriate SIP as having been executed at a FINRA TRF.

In addition to sending market data to the SIPs for consolidation, exchanges and FINRA TRFs are also permitted to sell access to the same transaction data directly to customers through proprietary data feeds. Importantly, the SEC has interpreted Rule 603 to require only that exchanges *transmit* data to the SIPs no later than they transmit data through their proprietary data feeds.[11] This implies that traders subscribing to a direct feed avoid the inevitable latency arising from the SIPs' obligation to consolidate and process transaction information before disseminating it.

---

[10] FINRA operates an Alternative Display Facility (the "FINRA ADF") through which non-exchange venues (such as an electronic communications network, or "ECN") might choose to disseminate quotations from their subscribers. At present, no venue disseminates any quotations through the FINRA ADF.

[11] See In re NYSE LLC, Exchange Act Release No. 34-67857, at 2 (Sept. 14, 2012). In adopting Reg NMS, the SEC similarly noted that while Rule 603 requires exchanges that offer proprietary feeds to do so on terms that are fair and reasonable and not unreasonably discriminatory, "Rule 603(a) will not require a market center to synchronize the delivery of its data to end-users with delivery of data by a Network processor to end-users." This SEC guidance accordingly permits subscribers of exchange data to *receive* this data before a SIP so long as the exchange *releases* the data to the subscriber no sooner than it does for the SIP.

To establish the magnitude of this delay, Table 1 provides processing times for trade and quote information disclosed by both SIPs from 2014 through the second quarter of 2016. For Tape A and B securities, the time between receipt of a transaction report by the NYSE SIP and its subsequent dissemination of that report averaged 410 microseconds for trades and 450 microseconds for quote updates. Processing times for Tape C securities were slightly higher at 700 microseconds and 750 microseconds, respectively.[12]

[Insert Table 1]

In addition to allowing exchanges to sell their direct feed data, the SEC also allows exchanges to sell co-location services. These services allow customers to place their computer servers in close physical proximity to the exchanges' matching engines to minimize the transit time of the exchanges' market data. For Tape A and B securities, co-location accordingly allows a trader to avoid the additional latency a transaction report experiences when traveling from a market center to the NYSE SIP in the NYSE's Mahwah, New Jersey datacenter (the same datacenter housing the NYSE's matching engine); for Tape C securities, it avoids the latency a report experiences when traveling to the Nasdaq SIP's processing platform in Carteret, New Jersey (the same datacenter housing Nasdaq's matching engine).

In light of widespread concerns about the advantages these direct feeds provide fast traders, SEC Chair Mary Jo White requested that the SIPs "incorporate a time stamp in their data feeds to facilitate greater transparency on the issue of data latency." (White, 2015) We use these new timestamps in the analyses below.

3. Data and Sample Selection

We obtain all trade and quote reports published by the two SIPs for the common stock of firms listed within the Dow Jones 30 as of August 1, 2015, which are made available through the NYSE Trade and Quote Daily Files. We focus initially on the Dow Jones 30 in light of popular claims that high frequency trading firms are "overwhelmingly interested in heavily traded" securities (Lewis, 2004: p. 115). Our

[12] The secular decline in processing-related latencies shown in Table 1 reflect several initiatives of both SIPs to improve processing speeds.

sample period commences with the full implementation of the new microsecond timestamps on August 6, 2015 (the first full day on which exchanges complied with the new reporting requirements) and ends on June 30, 2016.[13] To ensure that all quotes and trades occur during the trading day after the opening cross and before the closing auction, we subset the data to exclude quotes and trades occurring before 9:45:00 and after 15:44:59.999999. As noted in Holden and Jacobsen (2014), the NBBO file of the Daily TAQ file is incomplete; therefore, we manually calculate the NBBO for each security for each microsecond during our sample period using quote updates from the daily TAQ data and the standard Hasbrouck algorithm. In so doing, we restrict our analysis to those quotations that are eligible to establish an exchanges' best offer or best bid (i.e., quotation updates having a condition of A, B, H, O, R, W, or Y). Finally, for our latency analysis in Section 4, we exclude quote or trade records with missing venue timestamps or with venue timestamps that are subsequent to the SIP timestamp.[14] Imposing these conditions results in a core sample of 385,028,820 trades and 6,212,857,437 quote updates.

We next use these data to construct two versions of the NBBO that prevailed at the time of each trade in our sample. The first version calculates the NBBO using the timestamp showing the time (in microseconds) at which a SIP disseminated a quote update. This version reflects the NBBO that was available from the SIP at the moment of each trade; therefore, we designate it as the "SIP NBBO." The second version calculates an alternative NBBO using the new "Participant Timestamp," which shows the time (in microseconds) at which an exchange matching engine reported processing a quote update. This alternative version reflects the NBBO at the moment of each trade in a world with no processing or transmission latencies. Because it is derived directly from exchange data, we designate it the "Direct

[13] The implementation date for Tape C securities was July 27, 2015 and August 3, 2015 for Tape A and Tape B securities. However, the BATS Y exchange did not fully commence using the new timestamps until August 6, 2015. Prior to these amendments, SIP messages only carried timestamps marked in milliseconds that indicated when the processing of the messages by a SIP was completed, but not the time a venue processed a trade or a quote. The 2015 timestamp modifications also required clock synchronization among exchanges to ensure that timestamps are accurate within tolerances of 100 microseconds or less. See UTP Vendor Alert #2015 – 7: New Timestamp Definitions for July 2015 Release, available at https://www.nasdaqtrader.com/TraderNews.aspx?id=UTP2015-07.
[14] This sample selection rule excludes 55,226,095 quote updates (0.9% of all quotes), only 8 of which are due to missing venue timestamps, and 2,811,429 trade records (0.7% of all trades), none of which are due to missing venue timestamps. Our analyses in Section 5 include in the sample all quote and trade records with venue timestamps that are subsequent to the SIP timestamp, which are excluded in our latency analysis presented in Section 4.

NBBO." For both versions of the NBBO, we restrict attention to NBBOs as of 10:00am; because our record of quotes starts at 9:45am, this 15-minute "burn in" phase ensures that our first daily measure of the NBBO reflects the best quotes available across all exchanges.

Finally, we further exploit the new Participant Timestamps to match each trade to the SIP NBBO and Direct NBBO *that prevailed at the time the trade was executed*. We do so by assigning to each trade a SIP NBBO and a Direct NBBO based on the microsecond at which the trade was executed using the trade's Participant Timestamp. This approach differs from traditional approaches that assign the SIP NBBO to trades using only the trade's SIP timestamp, which was previously the only timestamp the SIP provided for a transaction. However, a trade's SIP timestamp may not reflect the SIP NBBO that prevailed at the time a venue actually executed the trade due to the transit and processing-related delays associated with the SIP's processing of trade reports. For similar reasons, relying on the SIP timestamp of a trade does not permit insight into the Direct NBBO that prevailed at the moment a venue executes a trade. Relying on the Participant Timestamp for trades thus permits a unique insight into how a broker or venue perceived the SIP NBBO and Direct NBBO at the very time they were seeking to price transactions, rather than the time at which the SIP processes the trade report.

We additionally classify trades as having been buy- or sell-side initiated using the Lee and Ready (1991) algorithm with no lag (see Bessembinder and Venkataraman, 2010). In so doing, we compare each trade's execution price to the SIP-NBBO assigned to the trade. This is the logical choice since our research question focuses on whether there is harm to traders on venues that price transactions at a potentially stale SIP NBBO. For all trades, we retain the SIP-generated timestamp on a trade report to permit analysis of trade reporting latencies.

Before turning to the results, we want to emphasize that the Direct NBBO is a construct rather than a direct measure. No trader has access to the Direct NBBO because of the physical distance between exchange matching engines. Nonetheless, the Direct NBBO provides an in-the-limit representation of the advantages of having access to exchanges' fastest direct feeds. In other words, to the extent that the need to receive and process quotes over direct data feeds diminishes the speed advantage of subscribing to these feeds over the SIP, our use of the Direct NBBO can be viewed as the outer maximum latency advantage a trader could expect by using direct feeds to construct the NBBO.[15] True direct feed data would presumably then show that the fraction of SIP-priced trades with equal SIP NBBO and Direct NBBO would be even higher than 97%. We note that Ding, Hanna, & Hendershott (2014), while focused only on a subset of exchanges, take advantage of direct measures.

4. SIP Reporting Latencies and Dislocations of the SIP NBBO and Direct NBBO

Before examining the economic benefits of trading at the Direct NBBO rather than the SIP NBBO, we first examine how the reporting latency of the SIP NBBO can result in dislocations between the SIP NBBO and the Direct NBBO. We define reporting latency as the difference between the timestamp of a transaction reported by a SIP (the "SIP Timestamp") and the Participant Timestamp, which is the time an exchange matching-engine or broker-dealer records a transaction as having occurred:

$$Latency = Timestamp_{SIP} - Timestamp_{Participant}$$

Across exchanges, we find a mean (median) reporting latency for quote updates of 1,130 (559) microseconds. Given our definition of reporting latency, this delay reflects both the time it takes for a quote update to travel from a reporting venue to the appropriate SIP as well as the time it takes for the SIP to process the message and place it on its multicast feed for distribution. As we are unaware of any prior work utilizing the new SIP timestamps, we provide an extended analysis of quote and trade reporting latencies by venue. Due space constraints, we present those analyses in an Internet Appendix.[16] As we document there, median reporting latencies for both quotes and trades are tightly correlated with the

[15] Even with need to receive and process quotes over direct data feeds, it is important to note that subscribers to direct data feeds have a structural advantage over SIP subscribers regardless of how fast the SIPs process the NBBO. This advantage arises from the fact that all Tape A and Tape B quote data must be managed, aggregated and disseminated by the NYSE SIP in Mahwah, while all Tape C information is processed by Nasdaq's SIP in Carteret. Thus, assuming just 50 microseconds of processing time by the NYSE SIP, a quote update made in Nasdaq that changes the NBBO for a Tape A security would not be known to a broker relying on the SIP NBBO who is located in Secaucus for 510 microseconds (290 microseconds from Carteret to Mahwah in fiber, 50 microseconds to process, and another 170 microseconds to send the data by fiber to the broker in Secaucus). Were the same broker to use direct feeds to construct the NBBO and assuming it also needed 50 microseconds to process the NBBO upon receipt of a quote update, the broker would see the NBBO change after just 188 microseconds from the quote update being made on Nasdaq (138 microseconds from Carteret to Secaucus and 50 microseconds to process).

[16] The Internet Appendix also analyzes the extent to which a venue's choice of the SIP NBBO rather than the Direct NBBO affects its trade execution statistics.

geographic distance between the reporting exchange and the relevant SIP.[17] This finding provides confidence that the participant timestamps accurately reflect the microsecond at which a quote occurs on a venue.

An inevitable consequence of these SIP reporting latencies is for the SIP NBBO to lag changes in the Direct NBBO. For instance, across all securities in our sample, the NBB from the SIP NBBO and that of the Direct NBBO differed on average 6,839 times per day. These differences—which, following Ding, Hanna, and Hendershott (2014), we refer to as "Dislocations"—ranged from a daily minimum of 206 for General Electric to a maximum of 138,644 for Apple.[18] However, as one would expect from the median reporting latency of quote updates, the duration of these dislocations was typically short-lived. Across all dislocations of the NBB, for example, the mean (median) duration was 1,001.6 (489) microseconds. A standard deviation of 567,350, however, highlights the existence of outliers. In Figure 1, we present a histogram of the duration of NBB dislocations which illustrates the thick-tailed nature of this distribution.[19]

[Insert Figure 1]

With regard to the size of these dislocations, mean and median dislocations for the NBB were \$0.0109 and \$0.01, with a 99th percentile of \$0.03. Dislocations of the NBO were similarly slight, having a mean, median and 99th percentile measure of \$0.0109, \$0.01, and \$0.03, respectively. These figures are consistent with the fact that securities in the sample often traded at or near penny spreads. Figure 2 shows the histogram of the magnitude of NBB dislocations, which emphasizes how tightly clustered around a penny these dislocations are. Penny dislocations are well over 90 percent of all dislocations. Dislocations

[17] As we note, the single exception is with respect to trade reports made by the National Stock Exchange which, in light of our findings, is investigating these reporting anomalies.
[18] The mean number of dislocations of the NBO was approximately 8,433, ranging from a minimum of 203 for GE to a maximum of 139,997 for Apple. We estimate NBBO dislocations starting at 10:00 am following a 15-minute burn-in phase.
[19] The duration of dislocations for the NBO are similar to those of the NBB. In the interest of space, we present results for the NBB only.

of two, three, and four pennies occur, but are rare. Dislocations of a nickel or above occur so infrequently they cannot be discerned in the graph.[20]

[Insert Figure 2]

5. Trading Losses from Trading at the SIP NBBO

In this section, we use the microsecond timestamps to investigate empirically the extent to which traders during our sample period could have been adversely affected by quote reporting latencies.

a. Estimated Losses to Liquidity Takers

We first examine claims that direct feed arbitrage can be used to harm retail traders whose market orders are filled by retail market makers having access to direct data feeds. As an example, suppose a direct feed showed the NBBO changing from $10.00 x $10.01 to $9.99 x $10.00, while the SIP's NBBO remained at $10.00 x $10.01. The SEC's recent investigation of Citadel presents evidence that Citadel' market-making desk might fill buy orders by selling to them at $10.01 (the stale NBO reflected in on the SIP feed) rather than at $10.00 (the NBO shown in its direct feed). Citadel could then cover by buying at $10.00 (the actual NBO), earning $0.01 of risk-free profit. However, as the investigation reveals, Citadel ended this strategy in January 2010 and it affected only 0.4 percent of Citadel's retail order flow, calling into question the extent to which retail traders remain subject to this form of direct feed arbitrage.[21]

Because our data does not discriminate between retail and non-retail orders, we estimate trading losses for all liquidity taking orders that are filled at a price equal to the SIP NBBO rather than the Direct NBBO. To do so, we exploit the fact that our dataset includes both the SIP NBBO as well as the Direct NBBO prevailing for every trade in our sample. This basic structure permits us to estimate investor losses in a two-step process. In step one, we identify those trades that match the SIP NBBO by defining an indicator variable "SIP Priced" that equals 1 when the trade price matches either the NBB or NBO as

[20] There are nonetheless some quite rare dislocations that are large in magnitude (e.g., over $1).
[21] The investigation also notes that when Citadel filled an order at a stale SIP price, it sought to cover in the market for less than 6.9% of the shares filled in this fashion. *See* In the Matter of Citadel Securities LLC, Jan. 13, 2017, available at https://www.sec.gov/litigation/admin/2017/33-10280.pdf.

reflected on the SIP NBBO, and equals 0 otherwise.[22] Trades that are "SIP priced" represent purchase and sale transactions that place the liquidity taker in the same position as if the venue priced the order using the SIP NBBO. Second, because trades priced at the SIP NBBO represent liquidity-taking orders that might have been at risk of direct feed arbitrage, we next compare how these SIP-priced trades would have been priced had they been priced at the Direct NBBO. We then measure whether a trade priced at the SIP NBBO rather than the Direct NBBO resulted in a loss or a profit for the trader placing the liquidity taking order.

In Table 2, we illustrate this two part process using 35 trades occurring in Apple, Inc. over a 15 millisecond period on November 13, 2015. The time set forth in the second column is the Participant Timestamp, which is the timestamp reported by the trading venue for when the trade occurred. We use the Participant Timestamp to place trades in chronological order. The Participant Timestamp gives us the ability to sort quotes and trades according to the moment they occurred, conferring knowledge of the actual quoting environment surrounding trades. For comparison, the third column presents the SIP Timestamp. Note that several pairs of trades, such as the fifth and sixth, are in chronological order according to the Participant Timestamp (reflecting the actual sequence in which they occurred) but not in chronological order according to the SIP Timestamp.

[Insert Table 2]

The fourth and fifth columns represent the NBB and NBO in effect at the time of the trade as reflected in the Direct NBBO, while the next two columns reflect the NBB and NBO as reflected in the SIP NBBO. As shown in the table, the trades commenced when the Direct and SIP NBBOs match at $113.37 x $113.38. At that time, however, the market data suggest an inter-market sweep order (ISO) to buy approximately 6,000 shares with a limit price of $113.39 was submitted to all exchanges sitting at the

[22] As noted in Bartlett & McCrary (2016), trading venues also frequently use the NBBO to price trades at its midpoint. However, because we require trade direction to evaluate a trade's profitability, we focus only on those trades priced at exactly the NBB or NBO which allows us to assign trading direction using the Lee-Ready (1991) algorithm.

NBO (BATS, Direct Edge A, Nasdaq, and the NYSE Arca).[23] Evidence of this order can be seen by the manner in which the first 30 trades (each marked with code "F" in column 8 for an ISO) sweep through these four exchanges (column 9), buying all shares on the venues that are offered for less than $113.40 (column 10). This order results in the Direct NBBO changing to $113.39 x $113.40 by 11:37:47.465000, at which time an apparently unrelated trade occurs in a non-exchange venue (Exchange Code="D"). At the time of this latter trade, however, the SIP NBBO now reflects a stale NBBO of $113.37 x $113.38. Following this non-exchange trade, the SIP NBBO updates to reflect the true NBBO so that the Direct NBBO and SIP NBBO match one another by the time of the last three trades.

For purposes of analyzing this sequence of trades, we focus on those trades whose price matched the SIP NBBO, identified in the column entitled "SIP Priced." Were these trades actually priced off the SIP, the SIP's delay has an economic effect only for the non-exchange trade (Trade #31, highlighted in bold) occurring immediately after the ISO order finished sweeping through the market and inducing a mismatch between the Direct NBBO and the SIP NBBO. Note, however, that we make no formal assumption about the data feeds actually used by a trading venue: Because we seek to investigate the economic costs of trading at stale SIP prices, our focus is on trades that are priced at the SIP NBBO regardless of whether it was because a venue actually used the SIP to price the trade or whether it used direct feeds that were simply too slow to update the venue's perception of the NBBO before the trade occurred. This focus on SIP-priced trades allows us to estimate the maximum economic consequence of trading at stale SIP prices even in an environment where traders and venues seek to minimize direct feed arbitrage.[24]

[23] ISO trades are those with an "F" in the trade condition code listed in column 8. An order marketed as an ISO is exempt from the Order Protection Rule of Reg. NMS, which prohibits a venue from filling an in-bound order if superior prices rest at other exchanges. As such, a trading venue receiving an inbound liquidity-taking ISO can fill it without checking other venues for better prices. However, the broker sending the ISO is responsible for sending simultaneous orders that sweep all venues with better prices. As such, ISO orders allow a trader to sweep through multiple levels of a venue's order book, as occurs in this example.

[24] This point is particularly relevant for trades in exchanges given that, as described in the Internet Appendix, many exchanges now use direct data feeds to price and route trades. In general, use of these data feeds is motivated by a desire to protect providers of displayed liquidity from having their quotes "picked off" by liquidity takers as the market moves against them. The existence of trades on these exchanges that are nevertheless priced at the stale SIP NBBO highlights how, notwithstanding the use of direct data feeds, exchanges can functionally trade at stale SIP prices. It is this inability of venues to receive instantaneous updates of the NBBO that has informed the use of

Based on the price of this SIP-priced trade, it appears to have been the result of a marketable buy order; therefore, the fact that the trade was filled at $113.38 (the stale NBO) rather than at $113.40 (the new NBO) allowed the originator of the order to save two pennies per share acquired, or $2.00 for the total order. Because we are testing for whether liquidity takers (such as the originator of this order) were harmed by trading at SIP prices, we record this trade as realizing negative "lost profits" of $2.00 (-$0.02 per share) because the liquidity taker *gained* rather than lost by trading at the SIP NBBO. The SIP NBBO and the Direct NBBO matched one another for all other trades, so the choice of NBBO had no effect on trade profitability for these other trades.

In Table 3, we generalize this type of analysis to our full sample of Dow Jones 30 trades. Panel A summarizes by exchange the percentage of trades that we classify as SIP Priced (weighted by transaction size) and their aggregate transaction value. We find approximately 75% of all shares traded in our sample were traded at prices that exactly match the SIP NBBO, representing transaction volume of approximately $3 trillion. Excluding shares traded in non-exchange venues, this percentage increases to approximately 88% and transaction volume declines to approximately $2.26 trillion.[25]

Panel B presents by trading venue the mean amount of lost profits per share that liquidity takers experienced by having their trades priced at the SIP NBBO rather than at the Direct NBBO. For each venue, means are size-weighted based on the number of shares traded. As noted previously, SEC investigations of direct feed arbitrage have focused primarily on whether retail market-makers are exploiting direct data feeds to the detriment of liquidity-taking retail orders. We therefore present results separately for trades reported to a FINRA TRF, which include all trades made by retail market-makers and trades made in dark pools. For completeness, however, we also present results for each exchange.[26]

"speed bumps" to deter HFT firms from exploiting the delay every venue experiences when updating its view of the NBBO.

[25] As discussed in the Internet Appendix, the Nasdaq SIP may have printed timestamps on messages approximately 200 microseconds before the Nasdaq SIP finished processing transaction reports in Tape C securities. All results in Table 3 are virtually identical when we conduct analyses after reducing the SIP timestamps in Tape C securities by 200 microseconds.

[26] Although concerns that retail traders are disadvantaged by direct feed data have focused on retail market-makers, an exchange's use of slow data feeds to calculate the NBBO can also matter for retail traders for at least two

Overall, Panel B indicates that liquidity-taking trades priced at the SIP NBBO had average lost profits of approximately -$0.0002 per share. As indicated in our Apple illustration, lost profits are defined as the Direct NBB minus SIP NBB for sell orders, and the SIP NBO minus the Direct NBO for buy orders. As such, these negative lost profits suggest that liquidity takers, on average, saved $0.0002 per share by having their trades priced at the SIP NBBO rather than at the Direct NBBO—the opposite of what would be expected if liquidity takers are systematically receiving inferior pricing due to SIP reporting latencies. Average savings for liquidity taking trades occurring on a non-exchange venue were slightly higher, having an average of $0.0003 per share.

To understand better the potential profits available from trading at these stale NBBO prices, Panel C expands the analysis to the full distribution of lost profits per share based on the number of trades in the sample. Given heightened concerns about direct feed arbitrage within dark pools, we present the distribution separately by exchange and non-exchange venues. The first column for each group provides the percentage of trades producing the specified amount of lost profits per share. The second and third columns provide the aggregate transaction value and the aggregate number of shares associated with these trades. The final column presents the aggregate lost profits for these trades (i.e., Total Shares x Lost Profits Per Share). While Panel B suggests liquidity-taking orders, on average, benefited when their traders were priced at the SIP NBBO, this final column permits insight into the maximum amount a retail market maker might have captured in our sample by exploiting this form of direct feed arbitrage.[27]

A notable finding reflected in Panel C concerns the extremely high frequency of trades having zero lost profits per share, which occurs when the SIP NBBO and Direct NBBO coincide at the time of a trade. As reflected in the distribution, a trade priced at the SIP NBBO rather than at the Direct NBBO had no

---

reasons. First, exchanges generally permit limit orders to be pegged to the bid, ask, or midpoint of the NBBO, as calculated by the exchange. Second, Rule 611 of Reg. NMS prohibits an exchange from trading-through a protected quotation; therefore, exchanges must route in-bound marketable orders to an exchange holding the NBBO if the exchange is unable to fill the order at a price that is at least as good as the NBBO. An exchange using slow data feeds might accordingly be at risk of filling in-bound market orders at stale NBBO prices by letting them hit pegged orders or by failing to route them to markets holding the actual NBBO.

[27] For ease of presentation, Panel C presents results with no adjustment to the Nasdaq SIP's timestamp; results using the adjusted timestamp are virtually identical to those shown in Panel C.

economic effect for approximately 97% of shares traded within our sample. As was apparent in our Apple illustration, it is only when the SIP and Direct NBBOs differ that the choice of NBBO matching can affect transaction prices. Accordingly, the high percentage of shares traded with zero lost profits reflects the simple fact that the SIP and Direct NBBO typically match one another.

For those trades that produced non-zero lost-profits per share, Panel C shows that nearly 90% of the trades (weighted by shares traded) produced better pricing for liquidity takers when the trade's price matched the SIP NBBO rather than the Direct NBBO. Specifically, among trades priced at the SIP NBBO, approximately 2.7% of shares traded on non-exchange venues and 2.2% of shares traded on exchanges had negative lost profits. Moreover, almost all of these instances cluster at -$0.01 lost profits per share.

In general, we attribute this distribution of lost profits to the fact that the NBBO will commonly change in response to serial buy (sell) orders so that late-arriving buy (sell) orders benefit from hitting stale quotes. For instance, in the Apple illustration above, the dark venue's delay in updating the NBBO to reflect the ISO buy order that induced a change in the Direct NBBO allowed the later-arriving non-exchange buy order to benefit by purchasing at the lower, stale NBO. This phenomenon also explains how even exchanges that use direct data feeds can be at risk of filling buy (sell) orders at bargain prices when marketable orders arrive as the market is in the process of moving higher (lower).

Reflecting this logic, Panel C highlights the remarkably low likelihood that a marketable order priced at the SIP NBBO received poorer pricing than it would have received had it been priced at the Direct NBBO. For non-exchange trades, Panel C indicates that, among trades priced at the SIP NBBO, just 0.2% of shares traded had a positive measure of lost profits. This estimate drops to 0.06% for exchange trades.

Finally, regardless of whether a liquidity taking trade benefited or suffered when priced at the SIP NBBO rather than the Direct NBBO, Panel C highlights that the overall economic significance of either result is manifestly small. For instance, summing aggregate lost profits for all trades where lost profits is negative amounted to $3.2 million for non-exchange trades and $8.4 million for exchange trades. As such, the maximum benefit to liquidity takers of having their trades priced at the SIP NBBO rather than the Direct NBBO was just $11.6 million across both types of venues, notwithstanding the fact that the aggregate value of SIP-priced trades was nearly $3 trillion. More importantly, the aggregate cost to liquidity taking orders for having their trades priced at the SIP NBBO rather than the Direct NBBO was even smaller. Specifically, the aggregate value of trades having a positive measure of lost profits was just $500,000—a figure which drops to $239,241 for trades occurring on non-exchange venues. In other words, even if all non-exchange trades represented fills by retail market-makers, the total profits that could be earned by filling retail trades in our sample at the SIP NBBO rather than the Direct NBBO would be less than $300,000 over our 11-month sample period.

b. Estimated Losses to Liquidity Providers

While the foregoing results suggest liquidity takers generally benefited when trades were priced at the SIP NBBO, the fact that there are two sides to every trade would suggest the reverse conclusion applied to liquidity providers. In our Apple illustration, for example, the buy order completed at the stale SIP NBO of $113.38 rather than at the new NBO of $113.40 meant the seller in the non-exchange venue who had posted the resting liquidity lost $0.02 per share by selling at the stale SIP NBO rather than selling at the Direct NBO. The mean measure of lost profits of approximately -$0.0002 per share accordingly highlights that to the extent trading at SIP prices adversely affects traders, these costs are more likely to be borne by liquidity providers than by liquidity takers.

Depending on the identity of the trader taking liquidity in these trades, this latter finding may point to the presence of the second form of direct feed arbitrage occurring in the market. In particular, our results have largely assumed that marketable orders reflect uninformed order flow, such as orders submitted by retail investors. Our basic finding that liquidity takers benefit from being priced at the SIP NBBO, however, is in principle also consistent with claims made by those such as IEX and the Chicago Stock Exchange that HFT firms routinely engage in a separate type of direct feed arbitrage. Under this alternative strategy, HFT firms submit marketable orders to "pick off" stale quotes posted by liquidity providers to earn risk-free profits.

The sequence of Apple trades in Table 2 provides an example of how such a strategy might work in practice. After having secured a "buy" trade at $113.38 (the stale SIP NBO) rather than at $113.40 (the new NBO), the trader submitting the buy order need only sell at the new NBB of $113.39 to realize an immediate, risk-free profit of $0.01 per share (excluding trading fees). Because the ensuing four trades each reflected "sell" transactions at this price, our Apple example—and the results in Table 3 more generally—may simply reflect the strategic use of marketable orders by HFT firms to pick off stale limit orders posted in venues that use slow quote feeds for pricing orders that are pegged to the NBBO. Moreover, the trader would have done so without incurring any market risk. And as highlighted in the introductory quotation to this article, it is the risk-free nature of direct feed arbitrage that has loomed large in debates surrounding the appropriateness of allowing exchanges to sell direct data feeds and the related emergence of "speed bumps" at exchanges such as IEX. This focus on the risk-free character of HFT profits arising from quote reporting latencies similarly informs theoretical accounts of the trading strategy. As summarized by Wah and Wellman (2013), "By anticipating [the] future NBBO, an HFT algorithm can capitalize on cross-market disparities before they are reflected in the public price quote, in effect jumping ahead of incoming orders to pocket a small but sure profit." The risk-free character of HFT firms "sniping" stale quotes also distinguishes direct feed arbitrage from other forms of high-speed trading that target stale limit orders to make risky bets on price trends.[28]

To explore the extent to which this form of direct feed arbitrage occurs in our sample, we leverage the new Participant Timestamp data and the fact that an HFT firm following such a strategy would need to make a pair of trades. To see how this works, consider trades immediately subsequent to those trades

[28] For instance, Harris and Schultz (1998) study "SOES bandits" who used Nasdaq's (now decommissioned) Small Order Execution System (SOES) to trade quickly in and out of positions. The SOES permitted immediate execution of trades; therefore, if market makers began updating quotes, a SOES bandit could use the system to enter a trade in hopes of hitting a slower market maker's quote that had yet to update. Harris and Schultz note that SOES bandits established positions "when they observe short-term trends" and close positions when "they feel a trend has run its course." In contrast to claims about HFT firms engaged in direct feed arbitrage, Harris and Schultz accordingly note that "SOES trading is risky" and that the strategy often resulted in sizeable losses.

where trading at the SIP NBBO yielded more favorable pricing for the liquidity-taking order than trading at the Direct NBBO—that is, where the trade produced a negative measure of lost profits. For each of these potential first-leg trades, suppose the trade originated from an HFT firm submitting to a venue an immediate-or-cancel buy or sell order after having observed a change in the Direct-NBBO. Because the trading strategy we test assumes risk-free profits from direct feed arbitrage, the success of this HFT strategy requires an off-setting second-leg trade—which one should be able to see in the data.

To execute this pairing strategy, we sort trades based on the Participant Timestamp and identify each potential first-leg trade based on whether it produced a negative measure of lost profits. We then search forward for a matching second-leg trade until a window of 1,000 microseconds from the first-leg trade timestamp has been exhausted. For a trade to match the first-leg trade, it must match both on the direction of the trade and the trade price. In particular, for first-leg buy orders, we require a matching second-leg trade to be a sell order at a price that is higher than the first-leg purchase price; conversely, for first-leg sell orders, we require a second-leg buy order at a price that is less than the first-leg sales price. We impose a 1,000 microsecond trading window following each first-leg trade to ensure there is sufficient time for a trader to receive a trade confirmation on the first-leg trade before placing the second-leg trade at either an exchange or non-exchange venue.[29]

Before presenting our results, it is worth emphasizing that this simple empirical strategy almost certainly over-estimates—potentially by a wide margin—the actual incidence of second-leg matches.

[29] We suspect a 1,000 microsecond trading window is most likely too generous for first-leg transactions occurring on stock exchanges. For instance, a trader subscribing to exchanges' fastest fiber optic data feeds and co-located at Nasdaq would receive a trade confirmation of a first-leg trade occurring at the NYSE (the furthest exchange from Nasdaq) in approximately 200 microseconds based on the reporting latencies presented in the Internet Appendix, allowing it to execute a second-leg trade even at the NYSE in approximately 400 microseconds from the time of the first-leg trade. Our choice of a 1,000 microsecond trading window is driven instead by the possibility that first-leg transactions occur on non-exchange venues. Given the execution risk assumed by a trader executing a first-leg trade, we assume an HFT firm choosing to use a non-exchange venue for the first-leg of this strategy would focus on those automated venues clustered within New Jersey's "Equity Triangle" that are capable of providing a trade confirmation with latencies comparable to those of the primary exchanges. However, as noted in the Internet Appendix, participant timestamps are recorded in milliseconds for non-exchange venues (rather than microseconds); therefore, imposing a 1,000 microsecond trading window for these trades has the effect of imposing a maximum window of between 501 microseconds (for a non-exchange trade that actually occurs at the 499th microsecond of a second) and 1,500 microseconds (for a non-exchange trade that actually occurs at the 500th microsecond of a second).

Among other things, for instance, our strategy disregards order size and transaction fees and focuses purely on identifying subsequent transactions that are priced higher (lower) than first-leg buy (sell) orders. Moreover, our approach seeks to identify matching second-leg trades independently for each first-leg trade, creating the possibility that the same second-leg trade can be matched to two different first-leg trades. Finally, our strategy also permits second-leg trades to occur in non-exchange venues, even though the absence of displayed liquidity in these venues would introduce considerable risk for the trading strategy.[30]

Even with this bias in favor of finding second-leg trades, the results of this analysis, which we present in Table 4, reveal an extremely low incidence of matches between first- and second-legs of this type of trading pairs. Table 4 stratifies the analysis between exchange and non-exchange venues because non-exchange venues may be more likely to price orders using the SIP NBBO. The results show that only 1.4% of all first-leg trades occurring in non-exchange venues can be matched to a second-leg trade within 1,000 microseconds. For first-leg trades occurring on exchanges, this percentage falls to less than 1%.

Given the strong bias our empirical strategy creates in favor of finding a second-leg matching trade, we interpret these results as confirming that the economic advantages for liquidity-takers of transacting at prices matching the SIP NBBO are unlikely to be the result of HFT firms seeking to exploit liquidity providers in venues that price transactions using slow quote data feeds or the SIPs. Our estimated upper bounds demonstrate that although anecdotal evidence may establish that these trading strategies exist, they are unlikely to be allocatively important in recent years for the Dow Jones 30. After all, if all negative lost profits amount to only $11.6 million in our sample, and roughly 0.8% of those trades might be part of an arbitrage play, arbitrage profits for the Dow Jones 30 are at most $110,000 over our sample period.

---

[30] Even for second-leg trades aimed at hitting an exchange's displayed liquidity, this strategy would appear to involve non-trivial execution risk. As noted in Fox, Glosten & Rauterberg (2015), an HFT firm attempting to profit from this form of direct feed arbitrage "must be able to transact against the new best quote before anyone else can."

c. Robustness of Results

A natural question is whether we can approximate how much money might be at stake from exploiting trades priced at stale SIP prices across the entire market over the course of a year, as opposed to just our sample period for just our sample stocks. Turning first to an annualized estimate of our sample, our sample period covers the period August 6, 2015 to June 30, 2016, or 228 trading days. As noted above, the aggregate value of trades having positive lost profits was approximately $500,000, which represents the total amount of money liquidity takers lost during this time period due to having their trades priced at the SIP NBBO rather than the Direct NBBO. There are 253 trading days in a calendar year, so we can estimate the annualized value of these trading losses for trades occurring on both exchange and non-exchange venues at approximately $555,000 (i.e., 253/228 x $500,000). Focusing on trades in non-exchange venues (which includes retail market makers), the annualized amount of positive lost profits would be just $265,474 (i.e., 253/228 x $239,241). This figure represents the aggregate amount liquidity takers in non-exchange venues lost by having their trades priced at the SIP NBBO rather than the Direct NBBO when trading in the Dow Jones 30 over the course of a year. To the extent these losses reflected profits to retail market makers, this amount would represent an outer estimate of the annual gross profits to retail market makers of exploiting quote reporting latencies given that non-exchange trades include trades made by retail market makers as well as those made in dark pools.

Focusing on the aggregate amount available to liquidity takers seeking to exploit stale quotes priced at the SIP NBBO, we estimate the annualized amount of risk-free profits from this strategy to be $122,061 (i.e., $110,000 x 253/228). Because this approach focuses on a two-legged trade to lock in risk-free profits, we can also calculate the gross benefit of picking off stale quotes at the SIP NBBO as simply the sum of all lost profits that have a negative value. These "gross profits" to liquidity takers can be understood as the pricing benefit of picking off a stale quote at the SIP NBBO whenever it saved the liquidity taker money relative to hitting an order at the Direct NBBO. This more generous approach for estimating the profitability of direct feed arbitrage amounted to $11.6 million for all trades in our sample, implying maximum annualized gross profits to liquidity takers of $12.8 million. However, this figure aggregates trades made both on and off exchanges, which may overstate the profitability of this strategy if traders on exchanges are less likely to use orders pegged to the NBBO. Focusing on those trades that occur on non-exchange venues, we find gross profits to liquidity takers in our sample of $3.2 million, implying annualized gross profits of $3.5 million.

A second consideration is that the Dow Jones 30 obviously cover only 30 stocks, and most of these stocks have quoted spreads at or near a penny. These small spreads diminish the profitability of these two forms of direct feed arbitrage to the extent average dislocations are likely to be at or near $0.01 per trade. Accordingly, to examine the robustness of our results, we re-estimate the annual profitability of direct feed arbitrage to liquidity providers and to liquidity takers for the entire trading market.

Because there are nearly 8,000 stocks observed over this time period, we estimate these figures using a computationally efficient process that leverages the fact that a small number of stocks account for a large volume of shares traded.[31] In particular, the top-traded 257 stocks correspond to half the trading volume during our sample period, and the top-traded 872 stocks correspond to three-quarters of the trading volume during our sample period. By measuring the profitability of direct feed arbitrage strategies within these two groups during our sample period, we can obtain market-wide estimates by scaling the actual measure by a factor of 2 and 4/3, respectively. Finally, we can convert these estimates to annualized figures by multiplying by 253/228, as noted previously.

We present the results in Table 5. As shown in the first column, there were approximately $19.7 trillion of trades during our sample period in the 257 most traded securities. Of these approximately $14.8 trillion were priced at the SIP NBBO. Focusing on the top 872 traded securities, these figures are $32.1 trillion and $23.5 trillion, respectively. Grossing up these figures by 2 and 4/3, respectively, and scaling to 12 months yields annual market-wide estimates of approximately $44-47 trillion in total securities traded, of which approximately $33-35 trillion were priced at the SIP NBBO.

---

[31] We ignore stocks with any suffixes on their trading symbol. Stocks with no suffixes correspond to 98.5% of trading volume over our study period, so this is of little consequence for the calculations we report here. Ignoring suffixes is computationally advantageous because of the database index structure.

Of these SIP-priced trades, we first calculate aggregate profits to liquidity providers engaged in direct feed arbitrage. Because gains to liquidity providers constitute losses to liquidity takers, we estimate these profits as the total trading loses for liquidity taking orders that were priced at the SIP NBBO rather than at the Direct NBBO for all trades within the top 257 and 872 stocks. In light of our definition of lost profits, we accomplish this by summing lost profits for every trade observed where lost profits was greater than zero. This has the effect of aggregating trades where the liquidity taker could have earned just a penny more per share by having the trade priced at the Direct NBBO with those trades where the dislocation between the SIP NBBO and Direct NBBO was far greater than a penny. (For instance, we observe occasional dislocations of more than $1.00 per share traded.)

As shown in Table 5, the maximum amount of profits to liquidity providers from this form of direct feed arbitrage for trades among the top 257 stocks is approximately $1.3 million for exchange trades and $1.3 million for non-exchange trades. Multiplying these figures by 2 and annualizing yields annual market-wide estimates of approximately $2.9 million in both cases. Applying this approach to the top 872 stocks, our estimates of these annual market-wide profits to liquidity providers is $4.4 million for exchange trades and $3.5 million for non-exchange trades. Because concerns about this form of direct feed arbitrage have focused on retail market makers, we view the estimates for non-exchange trades as particularly important for assessing the relevance of this strategy in today's markets. Moreover, using data from FINRA's Order Audit Trail System (OATS), Tuttle (2014) estimates that non-ATS trades reported to FINRA comprise 60% of non-exchange trades. To the extent this reflects the fraction of off-exchange trades at risk to this form of direct feed arbitrage, our market-wide estimate of annual profits decreases to a range of approximately $1.7 million to $2.1 million.

Turning to the maximum profits available to liquidity takers from picking-off stale quotes priced at the SIP NBBO, we present our estimates of the gross arbitrage profits from this strategy in the final two rows of Table 5. Again, given our definition of lost profits, we calculate the maximum available profits from this strategy by summing the total value of lost profits for every trade observed where lost profits was less than zero. In general, this approach assumes a single trader is capable of picking off *every* quote in the market that is priced at the stale SIP NBBO and doing so only in those cases where it resulted in a profit to the liquidity taking trader.

As shown in the table, the aggregate value of these gross profits to liquidity takers for the top 257 stocks during our sample period was approximately $57 million for exchange trades and $29.6 million for non-exchange trades. For the top 872 stocks, these figures were $116.8 million and $47.5 million, respectively. Grossing up all figures by 2 and 4/3, respectively, and annualizing yields annual market-wide estimates of total gross profits to liquidity takers of between $126 million and $173 million for exchange trades and $66 million to $70 million for non-exchange trades. Because concerns about HFT sniping of stale quotes have focused on dark pools that peg orders to the NBBO, these profit opportunities may be confined to the subset of non-exchange trades occurring in dark pools. Turning again to estimates from Tuttle (2014), estimating that approximately 40% of non-exchange trades occur in venues classified as alternative trading systems, we approximate maximum annual gross profits to liquidity takers from this form of Direct feed arbitrage as $26 million to $28 million. Again, these are gross profits to liquidity-takers. Therefore, to the extent a liquidity taker seeks risk-free profits from this strategy as discussed previously, this estimate would be further reduced by the costs of the off-setting trade, as well as trading commissions and exchange fees.

Finally, we note that Virtu Financial—a single firm utilizing HFT subject to Exchange Act reporting obligations—had in its 2016 fiscal year "communication and data processing" costs of $71 million and "brokerage, exchange and clearance fees" of $221 million on trading revenue of $665 million.[32] In combination with our findings in Table 5, these disclosures further confirm that the profitability of HFT is unrelated to these two forms of direct feed arbitrage, notwithstanding their prominence in contemporary debates concerning market structure.

[32] Virtu Financial, Inc, Form 10-K for the Fiscal Year Ending December 31, 2016, available at https://www.sec.gov/Archives/edgar/data/1592386/000155837017001698/virt-20161231x10k.htm.

6. Conclusion

In his 2014 book *Flash Boys*, Michael Lewis captured international attention through his depiction of an equity market that systematically favors high frequency traders over slower traders such as retail and institutional investors. Central to his critique was the sale to HFT firms of fast access to exchange quotation data, which enables them to predict changes in the SIP-generated NBBO that trading venues have historically used to price both market and limit orders. For retail market-making firms such as Citadel and KCG, this speed advantage means the possibility of filling in-bound market orders at NBBO prices they know to be stale. For other HFT firms, it means the possibility of picking off mispriced limit orders pegged to a SIP NBBO that has yet to reflect the prices these fast traders can foresee.

Using recently released data from the two SIPs, we present novel evidence regarding the merits of these claims in the current trading environment. Due in large part to the political fall-out from Lewis' narrative, these data now include the precise time at which a quote update or trade report was processed by the relevant SIP along with the time it actually occurred on a trading venue. As we show, the availability of this latter timestamp is especially important as it enables for the first time the reconstruction of the real-time sequencing of quote updates and trades across the entire market and, critically, how they relate to one another and to the NBBO published by the SIPs.

Exploiting these new data, we show that since the release of these timestamps in August 2015, liquidity-taking orders gain on average $0.0002 per share when priced at the SIP-reported NBBO rather than the Direct NBBO, which reflects the NBBO calculated in a world without any reporting latencies. This finding reflects the simple fact that dislocations between the SIP NBBO and Direct NBBO can occur in response to serial buy and sell orders, allowing late-arriving market orders to benefit if they are priced at an NBBO that has yet to reflect the new trading interest. To the extent this is the case, concerns about trading at venues that are slow to update the NBBO would accordingly seem more relevant for traders providing liquidity in venues that price limit orders pegged to the NBBO using the slower data feeds. Yet while these concerns are consistent with claims that HFT firms pick off mispriced limit orders in these venues, we find virtually no evidence of this strategic behavior using the new Participant Timestamp data.

In short, our findings reveal that transacting at prices that match the SIP-NBBO can benefit liquidity takers to the detriment of liquidity providers. However, the incidence of these gains and losses between these two forms of trading interest appears to be primarily a product of chance rather than of HFT design. Because our data commence in August 2015, we emphasize that these findings may very well reflect a new market environment in which the HFT strategies depicted in *Flash Boys* are less prevalent than in the past. Among other things, for instance, the increasing processing speed of the SIPs, enhanced regulatory scrutiny of HFT, and the emergence of venues such as IEX that shield traders from HFT trading may have simply made these SIP-oriented arbitrage strategies increasingly infeasible.

Finally, while our findings are consistent with the incentive of liquidity providers to invest in fast access to trading data to avoid being adversely selected through direct feed arbitrage, our results suggest these incentives play, at most, a subsidiary role in promoting the socially costly arms-race for trading speed described in Budish, Cramton & Shim (2015). Although our core sample includes approximately $3 trillion of trades priced at the SIP NBBO, liquidity providers trading at the SIP NBBO could have saved just $11.6 million in lost profits had they transacted at the Direct NBBO instead. To the extent traders participate in this arms race, the primary incentives today would accordingly appear to rest outside a desire either to exploit direct feed arbitrage or to avoid the costs of trading at stale SIP prices.

References

Bartlett, Robert, and Justin McCrary, 2016, Dark Trading at the Midpoint: Pricing Rules, SEC Enforcement Policy, and Latency Arbitrage, UC Berkeley School of Law Working Paper.

Bessembinder, Hendrik, and Kumar Venkataraman, 2010, Bid-Ask Spreads: Measuring Trade Execution Costs in Financial Markets, in "Encyclopedia of Quantitative Finance" (Rama Cont, Ed.), John Wiley & Sons.

Brogaard, Jonathan, Bjorn Hagstromer, Lars Norden, and Ryan Riordan, 2015, Trading fast and slow: Colocation and liquidity, *Review of Financial Studies* 28, 3407–43.

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

Ding, Shengwei, John Hanna, and Terrence Hendershott, 2014, How Slow is the NBBO? A Comparison with Direct Exchange Feeds, *Financial Review* 49, 313–332.

Egginton, Jared F, Bonnie F. Van Ness, and Robert A. Van Ness, 2016, Quote Stuffing, *Financial Management* 45, 583-608.

Foucault, Thierry, Johan Hombert, and Ioanid Rosu, 2016, News trading and speed, *Journal of Finance* 61, 335–82.

Fox, Merritt, Lawrence R. Glosten, and Gabriel Rauter, 2015, The New Stock Market: Sense and Nonsense, *Duke Law Journal* 65, 191-277.

Harris, Jeffrey H. and Paul H. Schulz, The Trading Profits of SOES Bandits, *Journal of Financial Economics*, 50, 39-62.

Hendershott, Terrence, and Pamela Moulton, 2011, Automation, speed, and stock market quality: The NYSE's hybrid, *Journal of Financial Markets* 14, 568–604.

Hoffmann, Peter, 2014, A dynamic limit order market with fast and slow traders, *Journal of Financial Economics* 113, 159–69.

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

Jovanovic, Boyan, and Albert Menkveld, 2012, Middlemen in limit order markets. Working Paper, University of Amsterdam.

Lee, Charles M.C. and Mark J. Ready, 1991, Inferring Trade Direction from Intraday Data, *Journal of Finance*, 46, 733-746.

Levinson, Charles, U.S. investigates market-making operations of Citadel, KCG, Reuters, May 10, 2016.

Lewis, Michael, 2014, *Flash Boys: A Wall Street Revolt*, W.W. Norton & Co.

Menkveld, Albert J. High Frequency Trading and the New Market Makers, *Journal of Financial Markets* 16, 712-740.

Menkveld, Albert J., and Marius Zoican, Forthcoming, Need for speed? Exchange Latency and Liquidity, Review of Financial Studies

O'Hara, Maureen and Mao Ye, 2011, Is market fragmentation harming market quality? *Journal of Financial Economics* 100, 459-474.

Tuttle, Laura, 2014, "Description of Non-ATS OTC Trading in National Market System Stocks," unpublished SEC research report, available at https://www.sec.gov/marketstructure/research/otc_trading_march_2014.pdf.

Wah, Elaine, and Michael Wellman, 2013, *Latency arbitrage, market fragmentation, and efficiency: A two-market model*, Proceedings of the fourteenth ACM Conference.

White, Mary Jo, 2015, Testimony on "Examining the SEC's Agenda, Operations and FY 2016 Budget Request", Before the United States House of Representatives Committee on Financial Services, available at https://www.sec.gov/news/testimony/2015-ts032415mjw.html.

Figure 1

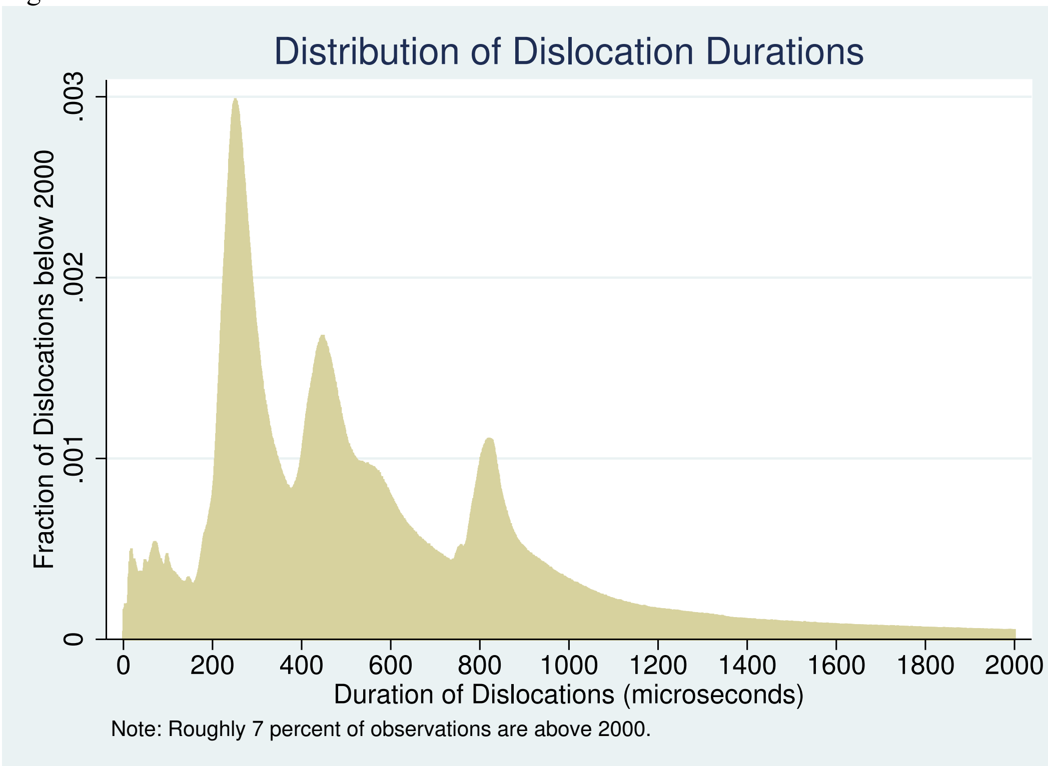


Figure 2

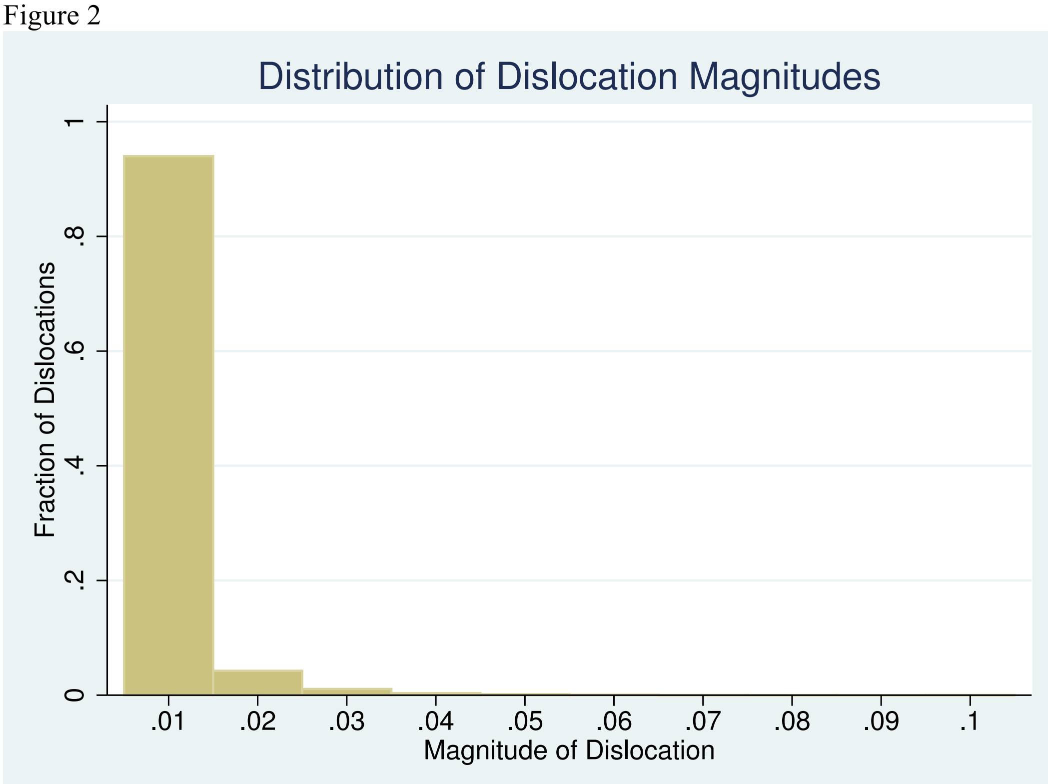


Table 1: SIP Processing Times

This table presents the processing times reported by the two SIPs for trade and quote data. Latencies are measured from the moment a trade or quote report is received by a SIP to the moment the SIP completes processing the record. Data for Tape A and Tape B securities can be found at https://www.nyse.com/publicdocs/ctaplan/notifications/trader-update/CTA%20SIP%202Q16%20Consolidated%20Data%20Operating%20Metrics%20Report%20(7-13-16%20Update).pdf. Data for Tape C securities can be found at http://www.utpplan.com/DOC/UTP%202015-Q4%20Stats%20with%20Processor%20Stats.pdf.

Panel A: SIP Processing Time for Trades

	Tape A&B Trade metrics						Tape C Trade metrics					
	Peak Messages per 100 Milliseconds (thousands)	Capacity Messages per 100 Milliseconds (thousands)	Capacity vs Peak Ratio	Average Latency	Median Latency	90th percentile latency	Peak Messages per 100 Milliseconds (thousands)	Capacity Messages per 100 Milliseconds (thousands)	Capacity vs Peak Ratio	Average Latency	Median Latency	90th percentile latency
1q14	21.80	60.00	2.75	0.51	n/a	0.71	19.30	39.40	2.04	1.32	1.25	1.67
2q14	23.50	60.00	2.55	0.51	n/a	0.66	20.50	39.40	1.92	0.82	0.54	0.74
3q14	22.70	65.00	2.86	0.51	n/a	0.66	17.60	48.50	2.76	0.59	0.49	0.68
4q14	24.20	65.00	2.69	0.45	n/a	0.60	19.40	48.50	2.50	0.59	0.49	0.67
1q15	22.10	70.00	3.17	0.45	n/a	0.59	20.10	68.70	3.42	0.53	0.45	0.60
2q15	31.80	70.00	2.20	0.34	n/a	0.43	22.80	132.80	5.82	0.54	0.46	0.62
3q15	27.10	75.00	2.77	0.32	0.24	0.41	16.10	132.80	8.25	0.58	0.47	0.64
4q15	43.70	75.00	1.72	0.31	0.24	0.41	18.60	132.80	7.14	0.62	0.47	0.66
1q16	42.40	86.00	2.03	0.33	0.25	0.43	19.40	132.80	6.85	0.77	0.49	0.76
2q16	37.40	96.00	2.57	0.34	0.24	0.45	28.20	132.80	4.71	0.63	0.48	0.68
mean	29.67	72.20	2.53	0.41	0.24	0.54	20.20	90.85	4.54	0.70	0.56	0.77

Panel B: SIP Processing Time for Quotes

	Tape A&B Trade metrics						Tape C Trade metrics					
	Peak Messages per 100 Milliseconds (thousands)	Capacity Messages per 100 Milliseconds (thousands)	Capacity vs Peak Ratio	Average Latency	Median Latency	90th percentile latency	Peak Messages per 100 Milliseconds (thousands)	Capacity Messages per 100 Milliseconds (thousands)	Capacity vs Peak Ratio	Average Latency	Median Latency	90th percentile latency
1q14	121.10	300.00	2.48	0.45	n/a	0.90	51.50	70.70	1.37	1.20	1.08	1.62
2q14	131.70	300.00	2.28	0.44	n/a	0.76	51.20	70.70	1.38	0.69	0.48	0.70
3q14	121.10	325.00	2.68	0.45	n/a	0.88	49.80	83.80	1.68	0.59	0.43	0.79
4q14	141.80	325.00	2.29	0.41	n/a	0.75	95.40	83.80	0.88	0.55	0.43	0.66
1q15	146.40	350.00	2.39	0.39	n/a	0.68	85.50	166.90	1.95	0.50	0.44	0.62
2q15	142.60	350.00	2.45	0.46	n/a	1.02	48.00	215.00	4.48	0.65	0.44	0.69
3q15	158.40	375.00	2.37	0.51	0.23	1.13	37.10	215.00	5.80	0.80	0.45	0.79
4q15	162.30	375.00	2.31	0.44	0.21	0.93	41.00	215.00	5.24	0.81	0.45	0.81
1q16	163.30	392.00	2.40	0.49	0.22	1.08	60.10	215.00	3.58	0.92	0.47	1.04
2q16	168.40	400.00	2.38	0.49	0.22	1.09	83.00	215.00	2.59	0.80	0.46	0.93
mean	145.71	349.20	2.40	0.45	0.22	0.92	60.26	155.09	2.90	0.75	0.51	0.87

Table 2: Apple Trades Ordered by Participant Timestamp, November 13, 2015

This Table illustrates how trades in the sample are matched to the prevailing SIP NBBO and Direct NBBO. *Participant Timestamp* is the time in microseconds at which a venue reports executing a trade. *SIP Timestamp* is the time the SIP placed the trade report on its multicast line for dissemination, which incorporates transit and SIP-processing latencies. The *NBB Direct* and *NBO Direct* are calculated using the Participant Timestamp for quote updates, which reflects the time an exchange matching engine processed a quote. The *NBB SIP* and *NBO SIP* are calculated using the traditional SIP Timestamp assigned to quotes, which reflects the time a SIP disseminated a quote update. The Direct NBBO is matched to each trade based on the Participant Timestamp of the trade and the Participant Timestamp of the Direct NBBO. The SIP NBBO is matched to each trade based on the Participant Timestamp of a trade and the SIP Timestamp of the SIP NBBO. *SIP Priced* is coded as 1 where Trade Price equals the SIP NBBO and zero otherwise. *Lost profits* are defined as the Direct NBB minus SIP NBB for sell orders, and the SIP NBO minus the Direct NBO for buy orders. Column 9 reports the exchange code used by the SIPs. Codes are: Z=BATS; K=Direct Edge A; Q=Nasdaq; P=NYSE Arca; D=FINRA TRF). See Sections 3 and 5 for additional details.

(1) Trade No.	(2) Participant Timestamp	(3) SIP Timestamp	(4) NBB Direct	(5) NBO Direct	(6) NBB SIP	(7) NBO SIP	(8) Trade Cond.	(9) Exch.	(10) Trade Price	(11) Trade Size	(12) Buy Order	(13) SIP Priced	(14) Lost Profits
1	11:37:47.464119	11:37:47.464616	113.37	113.38	113.37	113.38	@F	Z	113.38	2500	1	1	0
2	11:37:47.464119	11:37:47.464706	113.37	113.38	113.37	113.38	@F	Z	113.38	100	1	1	0
3	11:37:47.464119	11:37:47.464762	113.37	113.38	113.37	113.38	@F	Z	113.39	100	1	0	
4	11:37:47.464119	11:37:47.464792	113.37	113.38	113.37	113.38	@F	Z	113.39	100	1	0	
5	11:37:47.464119	11:37:47.464848	113.37	113.38	113.37	113.38	@F	Z	113.39	200	1	0	
6	11:37:47.464135	11:37:47.464743	113.37	113.38	113.37	113.38	@F	K	113.38	100	1	1	0
7	11:37:47.464135	11:37:47.464820	113.37	113.38	113.37	113.38	@F	K	113.38	200	1	1	0
8	11:37:47.464135	11:37:47.464861	113.37	113.38	113.37	113.38	@F	K	113.39	100	1	0	
9	11:37:47.464135	11:37:47.464889	113.37	113.38	113.37	113.38	@F	K	113.39	100	1	0	
10	11:37:47.464135	11:37:47.464916	113.37	113.38	113.37	113.38	@F	K	113.39	100	1	0	
11	11:37:47.464298	11:37:47.464673	113.37	113.38	113.37	113.38	@F	Q	113.38	100	1	1	0
12	11:37:47.464298	11:37:47.464727	113.37	113.38	113.37	113.38	@F	Q	113.38	100	1	1	0
13	11:37:47.464298	11:37:47.464777	113.37	113.38	113.37	113.38	@F	Q	113.38	100	1	1	0
14	11:37:47.464298	11:37:47.464806	113.37	113.38	113.37	113.38	@F	Q	113.38	100	1	1	0
15	11:37:47.464315	11:37:47.464834	113.37	113.38	113.37	113.38	@F	Q	113.38	200	1	1	0
16	11:37:47.464315	11:37:47.464875	113.37	113.38	113.37	113.38	@F	Q	113.39	100	1	0	
17	11:37:47.464315	11:37:47.464903	113.37	113.38	113.37	113.38	@F	Q	113.39	100	1	0	
18	11:37:47.464315	11:37:47.464929	113.37	113.38	113.37	113.38	@F	Q	113.39	100	1	0	
19	11:37:47.464315	11:37:47.464943	113.37	113.38	113.37	113.38	@F	Q	113.39	100	1	0	
20	11:37:47.464360	11:37:47.465298	113.37	113.38	113.37	113.38	@F	P	113.38	100	1	1	0
21	11:37:47.464360	11:37:47.465320	113.37	113.38	113.37	113.38	@F	P	113.38	100	1	1	0
22	11:37:47.464360	11:37:47.465337	113.37	113.38	113.37	113.38	@F I	P	113.38	73	1	1	0
23	11:37:47.464360	11:37:47.465352	113.37	113.38	113.37	113.38	@F	P	113.38	200	1	1	0
24	11:37:47.464397	11:37:47.465380	113.37	113.39	113.37	113.38	@F	P	113.39	500	1	0	
25	11:37:47.464397	11:37:47.465423	113.37	113.39	113.37	113.38	@F	P	113.39	100	1	0	
26	11:37:47.464397	11:37:47.465441	113.37	113.39	113.37	113.38	@F	P	113.39	100	1	0	
27	11:37:47.464397	11:37:47.465456	113.37	113.39	113.37	113.38	@F	P	113.39	100	1	0	
28	11:37:47.464397	11:37:47.465472	113.37	113.39	113.37	113.38	@F	P	113.39	100	1	0	
29	11:37:47.464397	11:37:47.465487	113.37	113.39	113.37	113.38	@F	P	113.39	100	1	0	
30	11:37:47.464397	11:37:47.465502	113.37	113.39	113.37	113.38	@F I	P	113.39	72	1	0	
31	**11:37:47.465000**	**11:37:47.467422**	**113.39**	**113.40**	**113.37**	**113.38**		**D**	**113.38**	**100**	**1**	**1**	**-0.02**
32	11:37:47.466000	11:37:47.511814	113.39	113.40	113.39	113.40		D	113.39	100	0	1	0
33	11:37:47.466018	11:37:47.466459	113.39	113.40	113.39	113.40		Z	113.39	100	0	1	0
34	11:37:47.475000	11:37:47.478795	113.39	113.40	113.39	113.40		D	113.40	245	1	1	0
35	11:37:47.479000	11:37:47.482618	113.39	113.40	113.39	113.40		D	113.40	805	1	1	0

Table 3: Gross Profits to Liquidity Takers and Liquidity Providers from Direct Feed Arbitrage

This table presents estimates of gains and losses to liquidity takers and to liquidity providers for transacting at prices equal to the SIP NBBO rather than the Direct NBBO across trades in the Dow Jones 30. Panel A presents the fraction of trades in the sample where the trade price matched the SIP NBBO and the aggregate transaction value of these SIP-priced trades. Panel B presents estimates of the mean amount of lost profit per share on all SIP-priced trades that liquidity takers experienced by having their trades priced at the SIP NBBO rather than at the Direct NBBO. Lost profits are defined as the Direct NBB minus SIP NBB for sell orders, and the SIP NBO minus the Direct NBO for buy orders. Therefore, positive measures reflect losses to liquidity takers and gains to liquidity providers; negative measures reflect gains to liquidity takers and losses to liquidity providers. Panel C presents the distribution of lost profits per share across all trades and provides estimates of the aggregate gross gains to liquidity takers and liquidity providers. In Panel B, robust standard errors are in parentheses. *** p<.01, ** p<.05, * p<.1

Panel A

Exchange	*% of Trades Matching SIP NBBO*	*Transaction Value of SIP-Priced Trades*
NYSE	92.66%	$519,437,823,595
NYSE MKT	72.10%	$2,538,795,453
NYSE Arca	90.15%	$297,152,874,525
Nasdaq OMX BX	88.85%	$64,661,284,230
NASDAQ OMX PSX	93.47%	$46,377,000,535
NASDAQ	90.17%	$551,383,078,922
BATS	88.06%	$256,043,483,771
BATS Y	90.28%	$118,575,973,632
Direct Edge A	93.03%	$87,342,210,121
Direct Edge X	92.54%	$306,040,680,052
Chicago Stock Exchange	10.07%	$7,591,528,235
National Stock Exchange	95.47%	$52,839,187
FINRA TRF	51.43%	$724,121,410,638
All venues:	75.33%	$2,981,318,982,896
All Exchanges:	88.53%	$2,257,197,572,258

Panel B

Exchange	Lost Profit Per Share
NYSE	-0.0002***
	(0.00001)
NYSE MKT	-0.0002***
	(0.00002)
NYSE Arca	-0.0002***
	(0.00001)
Nasdaq OMX BX	-0.0002***
	(0.00001)
NASDAQ OMX PSX	-0.0003***
	(0.00002)
NASDAQ	-0.0003***
	(0.00002)
BATS	-0.0003***
	(0.00003)
BATS Y	-0.0001***
	(0.00001)
Direct Edge A	-0.0002***
	(0.00001)
Direct Edge X	-0.0002***
	(0.00002)
Chicago	0.0000
	(0.00000)
NSX	-0.0002***
	(0.00004)
FINRA TRF	-0.0003***
	(0.00002)
All Venues	-0.0002***
	(0.00001)
All Exchanges	-0.0002***
	(0.00001)

Panel C

Lost Profit Per Share Traded	Non-Exchange Venues: Percent of Trades	Non-Exchange Venues: Transaction Value	Non-Exchange Venues: Total Shares	Non-Exchange Venues: Aggregate Lost Profits	Exchange Venues: Percent of Trades	Exchange Venues: Transaction Value	Exchange Venues: Total Shares	Exchange Venues: Aggregate Lost Profits
<-0.1	0.00%	$3,086,948	21,949	-$2,915	0.00%	$16,483,456	118,876	-$16,956
-0.1	0.00%	$1,804,518	12,016	-$1,202	0.00%	$6,926,969	46,829	-$4,683
-0.09	0.00%	$2,718,153	19,517	-$1,757	0.00%	$12,818,541	88,799	-$7,992
-0.08	0.00%	$5,728,906	38,128	-$3,050	0.00%	$19,869,549	138,041	-$11,043
-0.07	0.00%	$11,128,519	73,694	-$5,159	0.00%	$34,928,293	235,974	-$16,518
-0.06	0.00%	$14,553,168	99,575	-$5,974	0.00%	$66,380,139	453,721	-$27,223
-0.05	0.00%	$41,677,494	287,703	-$14,385	0.00%	$171,967,414	1,166,258	-$58,313
-0.04	0.01%	$81,642,779	583,097	-$23,324	0.01%	$356,400,340	2,502,360	-$100,094
-0.03	0.01%	$228,845,833	1,709,533	-$51,286	0.02%	$952,702,519	6,974,480	-$209,234
-0.02	0.06%	$852,621,997	7,377,069	-$147,541	0.08%	$3,322,465,360	27,654,830	-$553,097
-0.01	2.58%	$18,913,407,484	294,021,711	-$2,940,217	2.09%	$55,489,716,105	741,881,129	-$7,418,812
0	97.13%	$701,949,626,077	11,075,800,000	$0	97.73%	$2,194,566,936,719	34,625,100,000	$0
0.01	0.20%	$1,951,316,560	22,606,188	$226,062	0.06%	$2,051,673,295	21,170,418	$211,704
0.02	0.00%	$48,563,466	447,219	$8,944	0.00%	$93,492,699	820,440	$16,409
0.03	0.00%	$9,971,003	84,115	$2,523	0.00%	$19,294,316	154,342	$4,630
0.04	0.00%	$2,602,154	19,365	$775	0.00%	$8,020,152	65,402	$2,616
0.05	0.00%	$1,216,598	8,584	$429	0.00%	$3,341,600	27,251	$1,363
0.06	0.00%	$511,677	3,728	$224	0.00%	$2,138,077	16,712	$1,003
0.07	0.00%	$188,254	1,310	$92	0.00%	$480,210	4,018	$281
0.08	0.00%	$52,660	550	$44	0.00%	$460,131	3,875	$310
0.09	0.00%	$8,256	100	$9	0.00%	$326,037	2,785	$251
0.1	0.00%	$35,648	200	$20	0.00%	$249,769	2,085	$208
>0.1	0.00%	$102,486	800	$119	0.00%	$500,568	4,641	$874
Total:	100.00%	$724,121,410,638	11,403,216,151	-$2,957,569	100.00%	$2,257,197,572,258	35,428,633,266	-$8,184,316
Total Gains to Liquidity Takers (|Sum| of All Negative Lost Profits):				$3,196,810				$8,423,965
Total Gains to Liquidity Providers (Sum of All Positive Lost Profits):				$239,241				$239,649

Table 4: Estimates of Risk-Free Profits to Liquidity Takers

This table presents estimates of the incidence within the sample of a trading strategy where liquidity takers earn risk-free profits from picking-off stale quotes priced at the SIP NBBO. See Section 5 for details.

	Frequency of First-Leg Trades Having a Second-Leg Match	Std. Dev.	N
All Exchanges	0.007	0.002	9,201,335
Non-Exchanges	0.014	0.004	1,824,470
Combined	0.008	0.004	11,025,805

Table 5: Estimates of Annual Profit Opportunities from Direct Feed Arbitrage for the Entire Equities Market

This table provides estimates of the annual profitability of the two primary direct feed arbitrage strategies across all trades in every listed security. Estimates are obtained by extrapolating from all trades in the top-traded 257 securities and the top-traded 872 traded securities during our 10 month sample period. These securities represent half and three-quarters of all trading volume in all listed securities, respectively. All figures are obtained from TAQ data. *Total Trading Value* is measured as the dollar value of all trades in the relevant group of securities. *Value of All SIP-Priced Trades* is the value of these trades where the trade is priced at the SIP NBBO. *Maximum Available Profits to Liquidity Providers* is an estimate of the total profitability to market makers seeking to exploit direct feed arbitrage. It is calculated as the dollar value liquidity takers lost on all trades by having them priced at the SIP NBBO rather than the Direct NBBO, summed across all trades where liquidity takers received inferior pricing at the SIP NBBO. *Maximum Available Profits to Liquidity Takers* is an estimate of the gross profitability to liquidity takers from picking off stale quotes priced at the SIP NBBO. It is calculated as the dollar value liquidity providers lost on trades by having them priced at the SIP NBBO rather than the Direct NBBO, summed across all trades where liquidity providers received inferior pricing at the SIP NBBO. Annualized estimates for the entire market are obtained by multiplying measured values by 506/228 (i.e., 2 x 253/228) in the case of the top-traded 257 securities or 1012/684 (i.e., 4/3 x 253/228) in the case of the top-traded 872 securities.

	Trades in 257 Most Traded Securities		Trades in 872 Most Traded Securities	
	As Measured During Sample Period	Annualized Estimate of Whole Market	As Measured During Sample Period	Annualized Estimate of Whole Market
Total Trading Value	$19.7 trillion	$43.7 trillion	$32.1 trillion	$47.5 trillion
Value of All SIP-Priced Trades	$14.8 trillion	$32.8 trillion	$23.5 trillion	$34.8 trillion
Maximum Available Profits to Liquidity Providers - Exchange Trades	$1,320,305	$2,930,151	$2,947,107	$4,360,340
Maximum Available Profits to Liquidity Providers – Non-Exchange Trades	$1,304,467	$2,895,001	$2,398,773	$3,549,062
Maximum Available Profits to Liquidity Takers - Exchange Trades	$56,694,428	$125,821,845	$116,810,804	$172,825,342
Maximum Available Profits to Liquidity Takers - Non-Exchange Trades	$29,593,843	$65,677,564	$47,519,353	$70,306,411

Internet Appendix for "How Rigged Are Stock Markets?
Evidence from Microsecond Timestamps"

Section 1 of this Appendix analyzes SIP reporting latencies for trade and quote reports for all transactions involving Dow Jones 30 securities between August 6, 2015 (the first full day on which exchanges complied with the new SIP reporting requirements) and June 30, 2016. Section 2 of this Appendix analyzes the extent to which quote reporting latencies can affect a venue's trade execution statistics.

1. Estimating Trade and Quote Reporting Latencies

We define reporting latency as the difference between the SIP Timestamp and the Participant Timestamp. This definition resembles, but is distinct from, that used by Ding, Hanna, and Hendershott (2014). Those authors analyze the timestamp generated by a proprietary server located at BATS' trading center that receives transaction reports directly from select exchanges (BATS, Direct Edge, and Nasdaq) as well as from Nasdaq's SIP feed. Their definition of latency accordingly assesses the delay associated with receiving SIP market data relative to receiving market data from these select exchanges for a trader in Secaucus, New Jersey (the location of the BATS data center).

In contrast, our measure of latency represents the delay between the time a market center processes a transaction and the time when the appropriate SIP disseminates a report for the transaction. As such, it represents the delay created by: (a) the transit time from a market center to either the NYSE SIP or the Nasdaq SIP, as applicable, and (b) the time it takes for the relevant SIP to process and disseminate the transaction report. In this regard, it can be viewed as the floor latency experienced by all consumers of the SIP data, regardless of their location relative to the SIPs. Our measure also permits analysis of this latency across all market centers and for both NYSE- and Nasdaq-listed securities.

All timestamps are marked in microseconds; therefore, our measure of latency is formally in microseconds. We note, however, that the microsecond timestamps for trades by non-exchange venues are uniformly reflected as having occurred in intervals of 1,000 microseconds (i.e., 1 millisecond). We interpret this pattern as reflecting the fact that most non-exchange venues have continued to record transactions at the level of the millisecond.[1] Assuming this is the case, our measure of latency will accordingly be biased slightly higher for these trades to the extent the transaction did not occur at precisely the beginning of the reported millisecond. As we discuss below, the delay in transaction reporting for non-exchange trades is so large it could be measured in milliseconds—and hence microsecond precision is not necessary to get an accurate sense of latency for these transactions.

a. Institutional Background on Clock Synchronization.

Because our analysis relies on comparing timestamps imposed by two different data centers, an important preliminary issue to consider is clock synchronization. In particular, if the clock used by a SIP and the clock used by a market participant are not synchronized, our latency measure may be inaccurate. Not surprisingly, addressing similar clock synchronization concerns has also been central to the SEC's proposed Consolidated Audit Trail (CAT), which is designed to allow the reconstruction of all quote and trade activity across multiple market centers. In this subsection, we provide institutional details regarding why synchronization issues for this study, like the CAT more generally, are unlikely to be material in today's markets. Readers familiar with these issues from the CAT or otherwise are invited to skip to subsection (b) where we commence presentation of our empirical results.

Considering modern computer clock synchronization protocols, the scope for non-synchronized clocks in recent years is likely small. This is partly because of the demise of manual, mechanical time-stamping of transactions in favor of automated order-entry systems. For instance, Network Timing Protocol (NTP) clients have long been included in servers and personal computers, permitting computer clocks to be synchronized within milliseconds of the US national time standard, or UTC(NIST)

[1] FINRA has required since 2014 that firms report a trade's execution time in milliseconds when reporting trades to the FINRA facilities if the firm's system captures time in milliseconds. See FINRA Regulatory Notice 14-21 (May 2014), available at http://www.finra.org/sites/default/files/NoticeDocument/p506337.pdf. The new timestamp requirements permit FINRA to convert to microseconds any transaction times submitted in milliseconds by a FINRA member. See NasdaqTrader.com, UTP Vendor Alert #2015 - 7 : New Timestamp Definitions for July 2015 Release, available at https://www.nasdaqtrader.com/TraderNews.aspx?id=UTP2015-07. We assume that clocks record transaction time in milliseconds by rounding microseconds to milliseconds.

(Lombardi, 2000).[2] Alternative protocols such as IEE 1588 Precision Time Protocol (PTP) are also commonly available for more advanced servers, ensuring clock times are within nanoseconds of the UTC(NIST). In releasing the new microsecond time stamp specifications, the CTA, CQS and UTP accordingly required that exchanges use a clock synchronization methodology ensuring timestamp tolerances of 100 microseconds. In releasing its plan for the CAT, the SEC (2016) further reports that these tolerances apply to the two SIPs and that the absolute clock offset on exchanges averages just 36 microseconds.

Clock synchronization is potentially a greater issue for non-exchange venues and broker-dealers. In contrast to the 100 microsecond tolerance used by exchanges, in recent years FINRA required that all computer system clocks and mechanical time stamping devices of FINRA members be synchronized to within one second of the UTC(NIST).[3] In practice, however, brokers responsible for handling the largest share of trading volume appear to utilize clock syncing with much greater precision than this formal requirement. In anticipation of the CAT, for instance, FINRA recently adopted new Rule 4590 which reduces the drift tolerance for computer clocks that record transactions in OTC and NMS equity securities from one second to 50 milliseconds. In adopting the new standard, FINRA noted that firms accounting for 95 percent of reportable transactions to FINRA's Order Trail Audit System (OATS) already report events in milliseconds and comply with the 50 millisecond clock synchronization standard. Likewise, in responding to the proposed rule, dark pool operator IEX noted the standard could be further reduced below 50 milliseconds given the system capabilities of most FINRA firms, citing its own synchronization standard of one millisecond.

Indeed, for many of the most important FINRA members such as dark pool operators and broker-dealer internalizers, the emergence of co-location services has undoubtedly facilitated synchronization tolerances of far less than 50 milliseconds. For instance, firms such as IEX that are hosted by the Equinix

[2] National time standards are synchronized (essentially, averaged after removing outliers and consistent errors) to yield an international reference called Universal Coordinated Time. In the United States, the National Institute of Standards and Technology (NIST) maintains an atomic clock that serves as the country's primary time standard, or UTC(NIST). It generally tracks the UTC to within 5 nanoseconds.
[3] FINRA Rule 7430 applied through 2015.

NY4 datacenter in Secaucus, New Jersey (which also hosts the matching engines of BATS and Direct Edge) can utilize a service called "High Precision Time" offered through Perseus' Communications. The service allows synchronization with UTC(NIST) through both NTP and PTP protocols and promises "certified time stamps to subnanosecond accuracy."[4] In December 2014, Nasdaq announced it would be offering the same service to its customers at its U.S. datacenter in Carteret, N.J.[5] The 2015 Customer Guide for NYSE Euronext similarly offers four different connection protocols to the UTC(NIST) to ensure timestamp "accuracy on the order of nanoseconds."[6]

All of these factors reduce the likelihood that clock synchronization issues materially affect our latency measure, particularly for exchanges, but even for non-exchange venues and broker-dealers. However, as noted by Angel (2014), any non-zero synchronization tolerance and random variation surrounding it will introduce some degree of clock synchronization error when reconstructing market conditions using time-stamped records from multiple market centers. Consistent with these concerns, our data do reveal evidence of such residual noise in the form of transaction reports with negative latency. In particular, approximately 0.88% of quote updates and 0.72% of trade reports had a SIP Timestamp that *preceded* the time reported in the Participant Timestamp. Given that a transaction must be processed by a participant before it is even received by a SIP, these outcomes obviously represent physical impossibilities.

Close inspection of the data reveals that the majority of these reports resulted from clock synchronization issues at the NYSE Arca and the NYSE SIP.[7] For instance, between May 16, 2016 and June 6, 2016, more than 75% of the daily quote updates and trade reports emanating from the NYSE Arca had negative reporting latencies. These reports from Arca account for approximately 65% of quote

[4] See Equinix Press Release, Equinix is First to Offer Global Access to High Precision Time™ from Perseus Telecom, Sep. 10, 2014, available at http://www.prnewswire.com/news-releases/equinix-is-first-to-offer-global-access-to-high-precision-time-from-perseus-telecom-274588571.html.

[5] See Nasdaq Press Release, Perseus Selected by Nasdaq for Time Stamping Service at US Data Center, Dec. 3, 2014, available at http://www.prnewswire.com/news-releases/perseus-selected-by-nasdaq-for-time-stamping-service-at-us-data-center-300004156.html

[6] See Intercontinental Exchange, Infrastructure and Americas User Guide (Feb. 2015), available at www.nyxdata.com/doc/243267.

[7] According to NYSE Euronext, "The Arca issue… identified was due to a bug that was fixed." Personal communication between authors and NYSE officials, dated August 2, 2016.

updates having negative latencies and more than 77% of trade reports with negative latencies. Excluding the negative latencies appearing in these Arca reports for this two week period, negative latencies appeared in 0.31% of quote updates and 0.162% of trade reports. Of these, 99.953% of the quote updates and 99.917% of the trade reports were in Tape A securities and arose across all exchanges trading Tape A securities, indicating occasional clock syncing issues at the NYSE SIP.[8]

Evidence of these negative latencies within our sample highlight the potential of clock synchronization issues to arise even with the institutional structures described previously. Because we lack a record of the actual UTC(NIST) for each transaction report, we are unable to measure the extent to which clock syncing affects our measure outside of these negative latencies. However, in all analyses in subsection (b), we exclude from our sample any transaction report having a negative latency. As noted below, our resulting latency estimates generally reflect the institutional structure of the SIP-reporting regime, providing confidence that any residual clock synchronization issues do not materially bias our analyses.

b. SIP Reporting Latencies Across Trading Venues

As we are unaware of any prior work utilizing these new timestamps, we first report in this section some of the basic descriptive patterns of our latency measures. Table A.1 presents the mean, standard deviation, median, and 90th percentile measures of latency by trading venue according to where the transaction originated, both for quote updates (Panel A) and trade reports (Panel B).[9] Because securities within the Dow Jones are listed on both the NYSE and Nasdaq, we also separate transactions according to whether transaction reports were sent to the NYSE SIP (Tape A securities) or the Nasdaq SIP (Tape C securities).

---

[8] Analysis of negative latencies in the transaction reports for all NMS equity securities occurring on two randomly-chosen trading days reveals a clear dependency on the structure in which the NYSE SIP processes transaction reports. Under the technical specifications for the CTA Plan and the CQ Plan, transaction reports in Tape A securities and Tape B securities are processed separately across twenty-six multicast lines with each line processing approximately 250 securities according to its trading symbol. On both trading days, when a negative latency appeared for a security's quote update or trade report, negative latencies also appeared for the quote updates and trade reports of every other security assigned to the same multicast line before ceasing for all securities so assigned.

[9] As we show below, our latency measure exhibits a long and thick right-hand tail, implying that the median may be a better estimator of the center of the distribution than the mean.

In both panels, we group exchanges by the location of their matching engines to facilitate analysis of the role of transit time in explaining variation in reporting latencies. All three exchanges controlled by the NYSE (the NYSE, NYSE MKT, and NYSE Arca) are hosted in the NYSE's datacenter in Mahwah, New Jersey. The three exchanges owned by Nasdaq (Nasdaq, Nasdaq OMX BX, and Nasdaq OMX PSX) are hosted in Nasdaq's datacenter in Carteret, New Jersey. Five other exchanges are hosted in Equinix's NY4 and NY5 datacenters in Secaucus. This includes the four exchanges owned by BATS Global Markets—BATS Exchange, BATS Y, Direct Edge A, and Direct Edge X—as well as the matching engine of the Chicago Stock Exchange responsible for trades in all Dow Jones equity securities.[10] The Equinix facility also hosts the trading system for the National Stock Exchange (NSX), which recommenced trading on January 1, 2016 after ceasing operations in early 2015. However, we list NSX separately given its idiosyncratic reporting of Participant Timestamps, as described below.

In Figure A.1, we map the approximate location of these three datacenters to illustrate their physical distances from one another, as well as from the two SIPs. Because the NYSE runs the NYSE SIP, transaction reports in Tape A securities transmitted by an NYSE exchange need only travel the distance between the NYSE matching engine and the SIP processor within the Mahwah datacenter. Across all Tape A transaction reports, reporting latencies are accordingly the smallest for transactions occurring on an NYSE-owned venue. For instance, NYSE mean (median) quote update and trade report latencies are 690 (301) microseconds and 356 (298) microseconds, respectively.

In contrast, quote updates and trade reports in Tape A securities occurring on an exchange hosted in the Equinix Secaucus datacenters must travel approximately 21 miles before being processed by the NYSE SIP. Reports in Tape A securities occurring on a Nasdaq exchange must travel even further in light of the approximately 35 miles separating the Nasdaq datacenter from the NYSE's facility. These

[10] The Chicago Stock Exchange (CHX) also maintains a matching engine in Equinix's CH3 datacenter in Chicago. A "Matching Engine Committee" at CHX determines which of the two matching engines will handle transactions in securities that can be traded on the CHX. At present, only seventy securities are assigned to the Chicago matching engine; all others are matched in New Jersey, including all securities in our sample. See Chicago Stock Exchange, New Jersey Data Center Eligible Symbols (July 18, 2016), available at http://www.chx.com/market-data/nj-data-center/.

distances account for the larger latencies for Tape A securities for quote and trade reports arising from transactions on the Nasdaq- and Equinix-based matching engines. With regard to quote updates, for instance, median reporting latencies in Tape A securities for the three Nasdaq exchanges are approximately 900 microseconds, while those for the BATS exchanges range from 491 to 517 microseconds. Median reporting latency for quote updates occurring on the Chicago Stock Exchange (CHX) is slightly higher at 839 microseconds. Given that reporting latencies for the CHX are similarly higher than those of the BATS exchanges in Tape C transactions, we attribute these higher latencies to the superior network performance of the BATS-controlled exchanges. Reporting latencies for Tape C securities display the opposite patterns across the three groups of exchanges, consistent with the fact that the exchanges closest to Nasdaq's datacenter should have the shortest transit times.

The primary exception among exchanges to these geographic-centered patterns appears in the transaction reports for the National Stock Exchange. While median latency for quote updates are just slightly higher than for other Equinix-based exchanges, mean reporting latencies are considerably higher at approximately 18 milliseconds for Tape A securities and over 41 milliseconds for Tape C securities. As suggested by the extraordinarily large standard deviations reported in the table, these very high mean values reflect extreme outliers. Reporting latencies for trade reports are even more out of line with the latencies one would expect given NSX's geographic location relative to the NYSE- and Nasdaq-SIPs. For instance, mean (median) reporting latencies for trade reports at the NSX were nearly 53 milliseconds (52 milliseconds) for Tape A securities and 52 milliseconds (53 milliseconds) for Tape C securities. Notably, transaction reports can traverse the 700 miles from Chicago to the two SIPs in just 9 milliseconds.[11] These reporting latencies would accordingly appear to reflect either an idiosyncratic

[11] In unreported results, we calculated reporting latencies for the approximately seventy securities that continue to match on the CHX's matching engine in Chicago located in Equinix's CH3 datacenter. For Tape B securities, mean (median) reporting latencies were 9,139 (9,005) microseconds for quote updates and 9,402 (9,409) microseconds for trades. For Tape C securities, mean (median) reporting latencies were 8,853 (8,749) microseconds for quote updates and 8,190 (9,207) microseconds for trades. These latencies reflect the fact that both the NYSE SIP and the Nasdaq SIP are approximately 700 miles from the CHX matching engine in Chicago.

system for recording Participant Timestamps or extremely slow and inconsistent report processing at the NSX.[12]

Across exchanges, mean trade latency was generally lower than mean quote latency for securities on both tapes; however, this difference largely reflects the thick right-hand tail of the distribution of quote updates. Among trades on exchanges, for instance, the 90th percentile latency was roughly twice the size of the median; for quote updates, the 90th percentile latency was closer to four times the size of the median quote latency. Focusing on median latencies between trade reports and quote updates, the difference between trade and quote latency for exchanges falls considerably (Tape A Quotes=566 microseconds vs. Tape A Trades=604 microseconds; Tape C Quotes=551 microseconds vs. Tape C Trades=555 microseconds).

With the exception of the NSX, the distribution of latencies for both trade reports and quote updates for exchange transactions was unimodal with extreme kurtosis, highlighting both the strong clustering near the median as well as outliers. In Figure A.2 and Figure A.3, we present histograms of reporting latencies for all combinations of exchanges and tapes for both quote updates and trade reports.[13] As noted previously, the presence of outliers is particularly prominent within quote updates. Given the considerably greater number of quote updates during the trading day, the long right-hand tail for quote update latencies is consistent with concerns that the large volume of quote message traffic can occasionally overwhelm available network capacity (Nanex, 2014; Ye, Yao, and Gai 2012).

A striking exception to the unimodal distribution of reporting latencies appears in the reporting latency of non-exchange trade reports. In Figure A.4 and A.5 we present histograms of trade reporting

[12] Data in this study were shared with the NSX, which provided the following statement regarding these findings: "The NSX is aware of both the research into the SIP reported quote and trade reporting latencies and the variances reflected with respect to other exchanges regarding the trade reporting latencies. NSX will conduct its own review of the data to better understand the anomalies of the trade reporting latency times and look forward to working with the authors on their continuing market research."

[13] With the exception of the histogram for trade reports at the NSX, all histograms presented in Figures A.2 and A.3 are truncated at latencies of 4 milliseconds (approximately the 95th percentile of the overall distribution of latency) to facilitate visualization of distributional form. We truncate trade reports for the NSX at 100,000 microseconds given the large number of trades having latencies in excess of 4 milliseconds.

latencies for non-exchange trades in Tape A and Tape C securities, respectively.[14] As shown in both figures, the distribution of latency across the two tapes is both multi-modal and highly-skewed, resulting in mean and median latencies that are considerably higher than latencies for exchange trades. For Tape A securities, the mean (median) reporting latency is approximately 87 milliseconds (7.1 milliseconds); for Tape C, the mean (median) latency is approximately 101 milliseconds (7.0 milliseconds).

Two features of non-exchange trade reporting most likely account for the peculiar shape of these distributions. First, as noted previously, all non-exchange trades must be reported to one of two TRF facilities, thereby aggregating trades executed by automated wholesalers and dark pools as well as by smaller broker-dealers. While many dark pools and retail wholesalers are co-located at exchanges, smaller FINRA members may be located further away and may have slower network connections to the TRFs. Smaller members of FINRA may also have slower trade reporting protocols, particularly given the amount of time brokers are permitted to report trades under SEC and FINRA rule-making. For instance, Rule 601 of Reg. NMS simply states that brokers must report trades “promptly,” while FINRA requires trades to be reported to a TRF as soon as practicable, but no later than 10 seconds, following trade execution.[15] The gap between this slow formal requirement and the comparatively rapid actual implementation highlights the extent to which off-exchange reporting is today conducted through automated systems.

Additionally, regardless of the speed with which a broker reports a transaction to a TRF, reporting latencies for non-exchange trades are also increased by the double-legged nature of the TRF-reporting regime. For instance, a broker who chooses to report non-exchange trades to the NYSE TRF will first report trades in Tape C securities to the NYSE TRF in Mahwah, which will then report the trade to the Nasdaq SIP in Cartaret. For a broker based at the Nasdaq facility, such a process guarantees a reporting

[14] Because this is a data-rich environment, the structure of the density can be inferred from a histogram without resorting to smoothing choices and kernel density techniques (Silverman 1986).
[15] See FINRA, Regulatory Notice 14-16: Equity Trading Initiatives: OTC Trade Sequencing, available at https://www.finra.org/sites/default/files/notice_doc_file_ref/Notice_Regulatory_14-46.pdf

latency equal to at least the round-trip transit time between the Nasdaq and NYSE facilities before the Nasdaq SIP even begins processing the report.

Finally, our summary results also highlight what appears to be an inconsistency in time-stamping procedures between the NYSE SIP and the Nasdaq SIP. Evidence of this inconsistency appears in comparing the processing latency reported for trade and quote records in Table 1, and our latency measures in Table A.1. In Panel A of Table A.2, we set forth the median processing latency for quote updates in the second quarter of 2016 for the NYSE and Nasdaq SIPs from Table 1, as well as the median reporting latencies for quote updates in Tape A and Tape C securities for all exchanges other than the NSX.[16] We also present the difference between these two medians for each exchange, which represents an estimate of the median transit time experienced by a quote update for each exchange. To illustrate how this estimate compares with the theoretical minimum transit time, we present the time it takes for light to travel the same distance in a vacuum. Finally, we present the ratio of our estimated transit latency to this theoretical minimum. Panel B of Table A.2 does the same for trade reports.[17]

As shown in both panels, estimated transit times for quote updates and trade reports for Tape A securities range from approximately 2.5 to 8 times the theoretical minimum. These results are to be expected given that message signals travel slower in fiber optic cable than through a vacuum and must navigate additional networking frictions from an exchange matching engine to the NYSE SIP.

Results for Tape C securities, in contrast, reveal transit times that would appear to defy the laws of physics. For instance, the median reporting latency of approximately 523 microseconds for quote updates on BATS would mean that messages traveled the 16 miles from Equinix's facility in Secaucus to Nasdaq's datacenter in Carteret in approximately 63 microseconds—an astounding two-thirds the amount of time it would take for light to travel this same distance. For Tape C transactions on one of the three Nasdaq-owned exchanges, Table A.2 suggest the total time between the moment a transaction occurred

[16] We exclude the NSX given that, as noted previously, quote and trade reporting latencies do not appear to reflect the NSX's geographic location relative to the two SIPs.

[17] We omit theoretical minimum transit times where an exchange is located in the same facility as the SIP.

and the moment it was processed and disseminated by the Nasdaq SIP was *less* than the time it took the SIP to just process the report.

Given these findings, we sought to document the manner in which the two SIPs calculated message processing times and imposed the SIP timestamp. With regard to processing times, both SIPs define processing latency from the time a message is received from an exchange to the time it takes to place the message on the multicast feed for distribution.[18] The SIPs are less consistent, however, with respect to when they impose the SIP timestamps. For the NYSE SIP, the technical specifications of the CTS and the CQS were modified in connection with the roll-out of the new timestamps to make clear that the SIP timestamp "indicates the time that processing a message is completed." With respect to the Nasdaq SIP, technical specifications were also modified at this time; however, the definition of the "SIP Timestamp" was revised to state simply that it provides "the number of microseconds since midnight EST."

In light of these disclosures and our empirical results, we suspect Nasdaq's SIP may be placing its timestamp on a transaction report during its message processing routine, rather than at the conclusion of the routine as is done by the NYSE SIP.[19] Indeed, as shown in the final column of both panels in Table A.2, adding 200 microseconds to each of the median Tape C reporting latencies in Table A.2 would bring the estimated Tape C transit times from all exchanges in line with those of the Tape A latencies. We account for this possibility in our analyses in Section 5 of the paper by using two versions of "SIP Time" for Tape C trade and quote records. In one, we assume the SIP's timestamp represents the time the Nasdaq SIP placed the message on its multicast line; in the other, we add 200 microseconds to SIP Time. All results are the same regardless of the version utilized.

[18] See Financial Information Forum, https://www.fif.com/docs/fif_latency__member_input.ppt
[19] We rule out the possibility that the discrepancy arises from exchange clocks running faster than the Nasdaq SIP's since fast clocks on an exchange would affect reporting latencies of the NYSE SIP as well as the Nasdaq SIP.

2. Quote Reporting Latencies and Trade Execution Statistics

In addition to concerns about the adverse selection risks associated with direct feed arbitrage, a secondary concern with the availability of direct data feeds relates to the possibility that market centers might misreport their trade execution statistics using the SIP NBBO. Any divergence between the SIP and Direct NBBOs creates the possibility for conflicting trade execution measures on which market participants may rely when routing orders. For instance, in the series of trades in Apple illustrated in Table 2 in the main paper, the fact that a dark venue priced a buy order at the stale SIP NBO of $113.38 rather than the current NBO of $113.40 created two possible measures of price improvement. Using the SIP NBBO as the benchmark, the trade received zero price improvement—it was priced exactly at the SIP NBO of $113.38. However, using the Direct NBBO of $113.40, the trade would have received 2 cents of price improvement.

The challenge of dueling trade execution statistics is even more extreme for effective spreads, which all trading centers must disclose in their Rule 605 reports and which are routinely used as "the industry's acid-test quality measure" to rank trading venues (Alpert, 2015). Returning to the Apple example above, a venue that benchmarked trade execution to the SIP NBBO would record an effective spread of 0.01 for Trade #31; however, using the Direct NBBO, effective spreads for that trade would be 0.03.[20] Thus, even though the buyer paid two cents less than the Direct NBO of $113.40, the effective spread of 0.03 would suggest the trader received an *inferior* trade execution than if she had simply transacted at the Direct NBO.

This counterintuitive result stems from the basic arithmetic for calculating effective spreads, which seeks to infer price improvement based on the difference between a trade's price and the midpoint of the benchmark NBBO. This emphasis on a trade's distance from the midpoint of the benchmark NBBO can cause effective spreads to increase when a venue calculates them using an NBBO other than the one used

[20] Effective spreads are calculated as twice the difference between the trade price and the midpoint of the benchmark NBBO.

to price trades. This is especially true when an exchange handles orders that are to be priced by reference to the NBBO, such as orders pegged to the near, far, or midpoint of the NBBO.

To illustrate, consider a situation where the Direct NBBO is $10.00 x $10.01, but the SIP NBBO is $10.01 x $10.02. A venue that priced off the Direct NBBO and filled pegged orders at $10.00 (the NBB), $10.01 (the NBO), and $10.005 (the midpoint) would record effective spreads on these trades of 0.01, 0.01, and 0, respectively, if it used the Direct NBBO as its benchmark. If it used the SIP NBBO as its benchmark, these measures would be 0.03, 0.01, and 0.02. Conversely, a venue that priced pegged orders off the SIP NBBO and filled orders at $10.01 (the SIP NBB), $10.02 (the SIP NBO), and $10.015 (the SIP midpoint) would record effective spreads for these trades of 0.01, 0.01 and 0, respectively if it used the SIP NBBO as its benchmark and 0.03, 0.01 and 0.02 if it used the Direct NBBO instead. Situations could also arise in which effective spreads improve when a venue used a different NBBO benchmark than the one used to price trades, which simply underscores the potential for divergent NBBOs to create conflicting measures of trade execution for the same trade.[21]

At the same time, however, the extent to which rival NBBO benchmarks actually affect a trading venues' aggregate trade performance disclosures should be mitigated by the fact that dislocations between the SIP-generated NBBO and the NBBO generated by direct feeds are typically infrequent and, when they do occur, short-lived. As noted in the primary paper, for instance, pricing a trade at the SIP NBBO rather than at the Direct NBBO had no economic effect for approximately 97% of the trades within our sample of Dow Jones trades, while the mean (median) duration of NBBO dislocations was 1,001.6 (489) microseconds.

To estimate empirically how much the choice of NBBO benchmark affects venues' trade execution statistics, we calculate effective spreads for each trade in our sample using as our benchmark both the SIP

[21] These latter situations can occur, for example, if a venue attempts to fill a trade at the NBB or NBO of its benchmark NBBO, which happens to be the midpoint of the alternative NBBO. For instance, if the SIP NBBO stands at $10.00 x $10.01 and the Direct NBBO stands at $10.00 x $10.02, a venue that tries to fill a "buy" order at the SIP NBO of $10.01 would record an effective spread of 0.005 on the trade using the SIP NBBO as its benchmark for calculating effective spreads. However, using the Direct NBBO as the benchmark for calculating effective spreads would yield an effective spread for a trade at $10.01 of 0 since it happened to occur at the midpoint of the Direct NBBO.

NBBO and the Direct NBBO. Specifically, for each trade, we first calculate effective spreads using the prevailing SIP NBBO for the trade as our NBBO benchmark followed by using the prevailing Direct NBBO as our benchmark. Since we are interested in understanding the pricing of marketable orders at the NBBO, we exclude ISOs given that ISOs can be filled at prices worse than the NBBO. In all cases, we calculate effective spreads as a percentage of the quoted NBBO spread—generally known as the effective/quoted spread ratio (E/Q)—to account for variation in the size of the quoted spread for our sample securities.[22]

Table A.3 presents the results of this examination. In the first three rows, we present separately the analysis for the NYSE MKT, the Chicago Stock Exchange (CHX), and the NSX. We distinguish these three exchanges for institutional reasons: each disclosed using the SIP NBBO to price all un-priced, pegged orders during our sample period.[23] As discussed previously, this institutional choice can often favor the use of the SIP NBBO as the relevant E/Q benchmark to the extent these venues process a material volume of these un-priced, pegged orders.

Consistent with this prediction, using the SIP NBBO to calculate the E/Q ratio produces a more favorable trade execution measure for the NYSE MKT. Specifically, the E/Q ratio for non-ISO trades on the NYSE MKT was approximately 84.82% when calculated using the SIP NBBO as the benchmark, and 85.06% when calculated with the Direct NBBO. Results for the Chicago Stock Exchange and the National Stock Exchange, in contrast, were inconsistent with this prediction, most likely reflecting the lower volume of un-priced, pegged orders processed on these exchanges. For instance, in unreported results, we find that trades priced at the midpoint of the SIP NBBO constitute 6.5% of non-ISO trades on

[22] We base our calculation of the E/Q ratio on the methodology described by BATS Global Markets. See Execution Quality Definitions, available at https://batstrading.com/market_data/execution_quality/definitions/. In summary, this method restricts attention to trades that (a) occur when markets are neither locked nor crossed, and (b) that are within 10% of the NBBO. Because these conditions would imply analyzing a slightly different subsample of trades when the benchmark NBBO changes, we first construct a sample of trades meeting the above criteria using the SIP NBBO, second construct an analogous sample using the Direct NBBO, and finally use as our analysis sample the set of trades that are in both the first and second set.

[23] All U.S. stock exchanges have disclosed the market data sources used to price and route trades since 2015. These disclosures were prompted in a June 5, 2014 speech by SEC Chair Mary Jo White, where she requested equity exchanges to file with the Commission the data feeds used for purposes of order handling, order execution, and order routing.

the NYSE MKT, but only 0.66% on the Chicago Stock Exchange and 0% on the National Stock Exchange. Because these trades reflect the filling of midpoint peg orders (Bartlett & McCrary, 2016), this evidence would suggest these latter two venues process a lower volume of trades pegged to the NBBO.[24]

The subsequent nine rows present results for the remaining exchanges, which all disclose using direct feeds to price trades, as opposed to the SIP NBBO.[25] For these venues, Table A.3 indicates that using the SIP NBBO as the benchmark generally results in a worse E/Q ratio, while using the Direct NBBO produces a more favorable measure of trade execution costs.[26] For all exchanges showing a statistically significant difference in E/Q ratios, however, the difference between using the SIP NBBO and the Direct NBBO as a performance benchmark changes the measure by a relatively small amount. The effect ranges from a low of 0.01 percentage points for the Nasdaq PSX to a high of 1.85 percentage points for BATS X. These figures highlight the fact that, while the choice of NBBO benchmark affects effective spread calculations, the degree to which it does so is likely to be small in magnitude.

While we lack data on how individual non-exchange venues calculate effective spreads, the final row in Table A.3 provides an analysis analogous to that above for all non-ISO FINRA trades within our sample. Calculating the E/Q ratio using the SIP NBBO produces a ratio of approximately 73.71%—modestly lower than the 74.09% obtained using the Direct NBBO as a benchmark. Given that these venues are likely to price a large number of orders by reference to the NBBO (Bartlett & McCrary, 2016), this finding is consistent with claims that a substantial portion of non-exchange venues continue to price trades using the SIP NBBO. At the same time, the extraordinarily small difference between the two

[24] The trading rules for the CHX and the NXS also suggest these venues do not ordinarily rely on the NBBO to price trades. For instance, while the CHX permits "midpoint cross" orders, it does not support other pegged order types. *See* Chicago Stock Exchange, CHX Order Types Primer, available at http://www.chx.com/trading-information/order-types/. The NSX supports orders that are pegged to the near, far, and midpoint of the NBBO, however, all such orders are non-displayed. *See* National Stock Exchange, Select NSX Order Types and Modifiers, available at http://www.nsx.com/images/documents/publications/NSX_Order_Types_v3_0_1.pdf.

[25] We include the NYSE within this group, notwithstanding the fact that its SEC filings indicate that it uses the SIP to obtain top-of-the-book quote updates from other exchanges when pricing pegged orders. Given that the NYSE trades in only NYSE-listed securities, the fact that it also uses order data obtained directly from its own matching engine has the practical effect of giving it a direct feed to a critical source of quote updates for Tape A securities.

[26] The single exception is for Nasdaq PSX, which accounts for less than 1% of all trades.

calculations further underscores the conclusion that the short-lived nature of dislocations between the SIP and Direct NBBOs greatly diminishes the potential for a venue's choice of NBBO to have a meaningful effect on its published effective spreads.[27]

[27] That NBBO dislocations can matter at all, however, nevertheless underscores the limitations of the prevailing system governing order execution disclosures. Initially implemented as Rule 11Ac1-5 in 2001, Rule 605 makes no mention of *which* NBBO to utilize as a performance benchmark when calculating order execution statistics; however, subsequent SEC guidance suggests market centers should utilize data from the SIP when complying with the rule. Given the large number of venues using direct feeds to price transactions, we believe any such endorsement of the SIP NBBO in Rule 605 reporting has the potential to bias trade performance metrics, as shown in Table 6. At the same time, permitting venues to choose their NBBO benchmark (as appears to be tolerated by the SEC) complicates interpretation of a venue's order execution information without disclosure of this information. For instance, certain venues have expressly declined to follow the SEC's guidance to use the SIP NBBO in calculating their Rule 605 reports, opting instead to calculate trade performance statistics using the same market data used to price transactions. See, e.g., IEX ATS Rule 605 Disclosure of Order Execution Information, available at http://50.116.60.129/regulation/605/. Requiring venues to disclose the NBBO benchmark used for calculating their performance metrics would represent a logical modification of Rule 605 given the divergent use of market data among trading centers.

References

Alpert, Bill, Who Makes Money on Your Stocks, *Barrons*, Feb. 27, 2015.

Angel, James, 2014, When Finance Meets Physics: The Impact of the Speed of Light on Financial Markets and their Regulation, *Financial Review* 49, 271-281.

Bartlett, Robert, and Justin McCrary, 2016, Dark Trading at the Midpoint: Pricing Rules, SEC Enforcement Policy, and Latency Arbitrage, UC Berkeley School of Law Working Paper.

Ding, Shengwei, John Hanna, and Terrence Hendershott, 2014, How Slow is the NBBO? A Comparison with Direct Exchange Feeds, *Financial Review* 49, 313–332.

Lombardi Michael, 2000, Computer time synchronization, National Institute of Standards and Technology working paper, available at: http://www.tf.nist.gov/service/pdf/computertime.pdf

Nanex, The Quote Stuffing Trading Strategy, Aug. 15, 2014, available at http://www.nanex.net/aqck2/4670.html

Securities and Exchange Commission, 2016, Joint Industry Plan; Notice of Filing of the National Market System Plan Governing the Consolidated Audit Trail, available at https://www.sec.gov/rules/sro/nms/2016/34-77724.pdf.

Silverman, Bernard, 1986, *Density Estimation for Statistics and Data Analysis*. Boca Raton: Chapman & Hall/CRC Press.

Ye, Mao, Chen Yao, and Jiading Gai, 2012, The externalities of high frequency trading, Working paper, University of Illinois at Urbana-Champaign

Table A.1: Quote and Trade Latencies

This table presents statistics regarding the latency of trade and quote reports for the Dow Jones 30 during our sample period. Panel A reports latencies for quote updates according to the venue making the quote update. Panel B reports latencies for trade reports based on the trading venue where the transaction occurred. Latencies are measured as the time (in microseconds) between the microsecond when the transaction occurs on the venue and the microsecond at which either the NYSE SIP (the SIP for Tape A securities) or the Nasdaq SIP (the SIP for Tape C securities) completes processing the transaction report.

Panel A: Quote Updates

	Tape A Securities					Tape C Securities				
Venue	*N*	*mean*	*sd*	*median*	*90p*	*N*	*mean*	*sd*	*median*	*90p*
NYSE	1,362,744,432	690	1,579	301	1,297	-	-	-	-	-
NYSE MKT	-	-	-	-	-	27,706,440	1,304	2,554	937	1,476
NYSE Arca	403,251,799	783	8,394	302	1,513	294,642,117	1,547	3,207	977	2,245
Nasdaq OMX BX	221,620,468	1,799	10,441	877	2,745	62,733,627	762	2,926	325	1,018
NASDAQ OMX PSX	258,795,046	1,972	12,730	923	3,297	87,393,187	886	3,335	367	1,246
NASDAQ	793,107,717	1,587	10,066	933	2,551	419,195,751	1,194	10,353	404	2,017
BATS	590,111,028	1,255	2,679	507	2,630	242,481,473	999	3,305	523	1,251
BATS Y	355,567,830	916	2,029	486	1,609	100,514,420	974	3,354	510	1,202
Direct Edge A	223,325,479	829	1,776	491	1,406	86,102,843	1,065	3,620	529	1,384
Direct Edge X	442,063,443	1,147	2,627	517	2,238	239,827,518	1,017	3,324	526	1,274
Chicago Stock Exchange	827,450	1,019	5,418	839	1,120	209,724	849	2,130	722	994
National Stock Exchange	529,478	18,073	3,657,389	1,228	2,073	106,167	41,176	6,307,078	962	1,992
All:	4,651,944,170	1,116	39,536	566	2,015	1,560,913,267	1,152	52,368	551	1,697

Panel B: Trades

	Tape A Securities					Tape C Securities				
Venue	*N*	*mean*	*sd*	*median*	*90p*	*N*	*mean*	*sd*	*median*	*90p*
NYSE	41,035,340	356	352	298	410	-	-	-	-	-
NYSE MKT	-	-	-	-	-	590,311	1,166	3,066	954	1,161
NYSE Arca	24,039,351	573	7,136	368	846	15,719,357	1,309	4,940	992	1,408
Nasdaq OMX BX	7,913,775	1,153	9,425	893	1,479	3,432,209	495	1,886	334	692
NASDAQ OMX PSX	3,478,058	1,125	6,511	903	1,575	2,656,520	503	1,869	345	746
NASDAQ	53,492,822	1,218	5,148	957	1,508	23,557,237	639	7,744	375	943
BATS	26,033,986	773	898	585	1,077	14,344,829	1,154	5,178	559	1,114
BATS Y	15,244,556	704	680	565	902	6,994,690	788	2,583	528	953
Direct Edge A	10,100,225	682	455	576	871	5,053,145	712	1,862	547	881
Direct Edge X	22,368,427	728	587	590	984	15,308,697	855	2,733	575	1,029
Chicago Stock Exchange	10,811	1,262	627	1,168	1,470	3,676	1,101	917	1,010	1,255
National Stock Exchange	11,886	52,824	29,633	52,447	94,009	4,705	52,473	29,350	53,255	92,826
FINRA TRF	64,940,748	86,979	2,311,033	7,149	115,260	28,693,459	101,277	371,501	6,982	176,022
All	268,669,985	21,624	1,136,806	849	8,344	116,358,835	25,648	189,542	717	10,226
All (excluding FINRA)	203,729,237	791	4,203	604	1,155	87,665,376	894	5,243	555	1,157

Table A.2: Estimates of Transit Latencies Relative to Theoretical Minimum Transit Times

This table presents estimates of the time it takes (in microseconds) for a transaction report to traverse from a reporting venue to either the NYSE SIP (for transactions in Tape A securities) or the Nasdaq SIP (for transactions in Tape C securities). Panel A estimates transit times for quote updates. Panel B estimates transit times for trade reports. *Median SIP Processing Time* refers to the median time it takes for the relevant SIP to process quote updates or trade reports (as appropriate) during the second quarter of 2016 based on data in Table 1. *Median Reporting Latency* is the median reporting latency for quotes and trades for the relevant trading venue as reported in Table A.1. *Estimated Transit Time* is the difference between Median Reporting Latency and Median SIP Processing Time and reflects an estimate of the time it takes for a transaction report to travel between the reporting venue and the appropriate SIP. *Theoretical Minimum Transit Time* is the time (in microseconds) it takes for light to travel the same distance in a vacuum.

Panel A: Quotes

	Tape A Securities					Tape C Securities					
Exchange	*Median SIP Processing Time*	*Median Reporting Latency*	*Difference (Estimated Transit Time)*	*Theoretical Minimum Transit Time*	*Estimated Transit Time / Theoretical Minimum*	*Median SIP Processing Time*	*Median Reporting Latency*	*Difference (Estimated Transit Time)*	*Theoretical Minimum Transit Time*	*Estimated Transit Time / Theoretical Minimum*	*Adjusted Ratio (+200 ms Reporting Latency)*
NYSE	220	301	81	-	-	460	-	-	-	-	-
NYSE MKT	220	-	-	-	-	460	937	477	188	2.5	3.6
NYSE Arca	220	302	82	-	-	460	977	517	188	2.8	3.8
Nasdaq OMX	220	877	657	188	3.5	460	325	-135	-	-	-
NASDAQ OMX	220	923	703	188	3.7	460	367	-93	-	-	-
NASDAQ	220	933	713	188	3.8	460	404	-56	-	-	-
BATS	220	507	287	113	2.5	460	523	63	86	0.7	3.1
BATS Y	220	486	266	113	2.4	460	510	50	86	0.6	2.9
Direct Edge A	220	491	271	113	2.4	460	529	69	86	0.8	3.1
Direct Edge X	220	517	297	113	2.6	460	526	66	86	0.8	3.1
Chicago	220	839	619	113	5.5	460	722	262	86	3.0	5.4

Panel B: Trades

	Tape A Securities					Tape C Securities					
Exchange	*Median SIP Processing Time*	*Median Reporting Latency*	*Difference (Estimated Transit Time)*	*Theoretical Minimum Transit Time*	*Estimated Transit Time / Theoretical Minimum*	*Median SIP Processing Time*	*Median Reporting Latency*	*Difference (Estimated Transit Time)*	*Theoretical Minimum Transit Time*	*Estimated Transit Time / Theoretical Minimum*	*Adjusted Ratio (+200 ms Reporting Latency)*
NYSE	240	298	58	-	-	480	-	-	-	-	-
NYSE MKT	240	-	-	-	-	480	954	474	188	2.5	3.6
NYSE Arca	240	368	128	-	-	480	992	512	188	2.7	3.8
Nasdaq OMX	240	893	653	188	3.5	480	334	-146	-	-	-
NASDAQ OMX	240	903	663	188	3.5	480	345	-135	-	-	-
NASDAQ	240	957	717	188	3.8	480	375	-105	-	-	-
BATS	240	585	345	113	3.1	480	559	79	86	0.9	3.2
BATS Y	240	565	325	113	2.9	480	528	48	86	0.6	2.9
Direct Edge A	240	576	336	113	3.0	480	547	67	86	0.8	3.1
Direct Edge X	240	590	350	113	3.1	480	575	95	86	1.1	3.4
Chicago	240	1168	928	113	8.2	480	1010	530	86	6.2	8.5

Table A.3: Effect of Using the SIP NBBO vs. Direct NBBO on Effective Spreads

This table shows the consequence of using the SIP NBBO rather than the Direct NBBO as the benchmark for calculating effective spreads for sample trades. *E/Q Ratio SIP NBBO As Benchmark* refers to the average effective/quoted spread ratios for trades on the relevant trading venue using the SIP NBBO as the pricing benchmark. *E/Q Ratio Direct NBBO As Benchmark* refers to the average effective/quoted spread ratios for trades on the relevant trading venue using the Direct NBBO as the pricing benchmark. *** p<.01, ** p<.05, * p<.1

Venue:	E/Q Ratio SIP NBBO As Benchmark	E/Q Ratio Direct NBBO As Benchmark	Difference	N
NYSE MKT	0.8482	0.8506	-0.0024***	137,825
Chicago Stock Exchange	15.6757	15.6728	0.0029	11,515
National Stock Exchange	0.9986	0.9943	0.0043***	1,337
NYSE	0.9115	0.8985	0.0130***	17,379,603
NYSE Arca	0.8933	0.8860	0.0073***	14,579,292
Nasdaq	0.9114	0.8957	0.0157***	34,388,463
Nasdaq BSX	0.9324	0.9300	0.0024***	6,727,836
Nasdaq PSX	0.9103	0.9104	-0.0001***	2,197,674
BATS X	0.8825	0.8639	0.0185***	17,470,674
BATSY	0.9691	0.9572	0.0119***	14,079,200
DirectEdge A	0.9714	0.9613	0.0101***	9,377,659
DirectEdge J	0.9536	0.9379	0.0157***	16,640,596
FINRA	0.7371	0.7409	-0.0038***	92,262,303

Figure A.1: Geographic Distance Between Datacenters

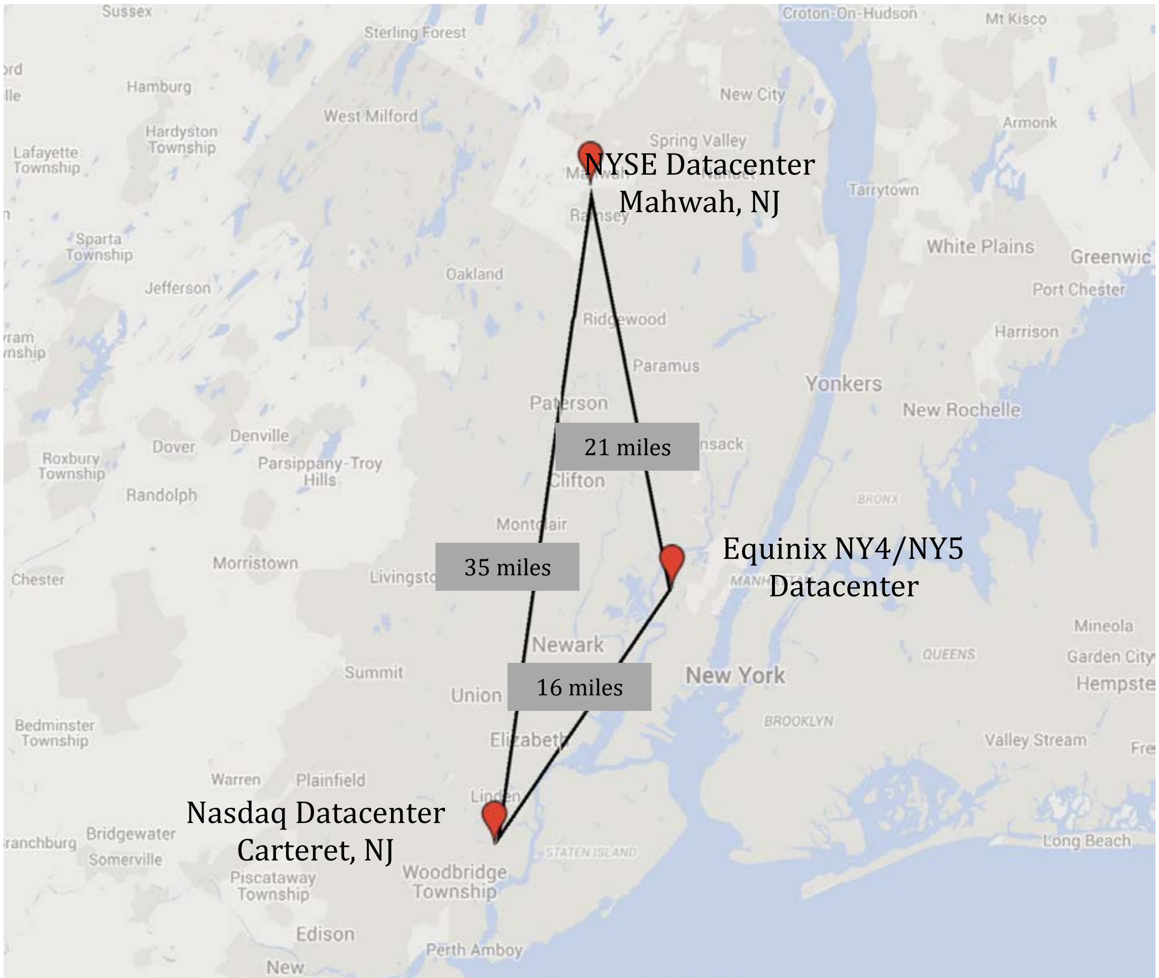
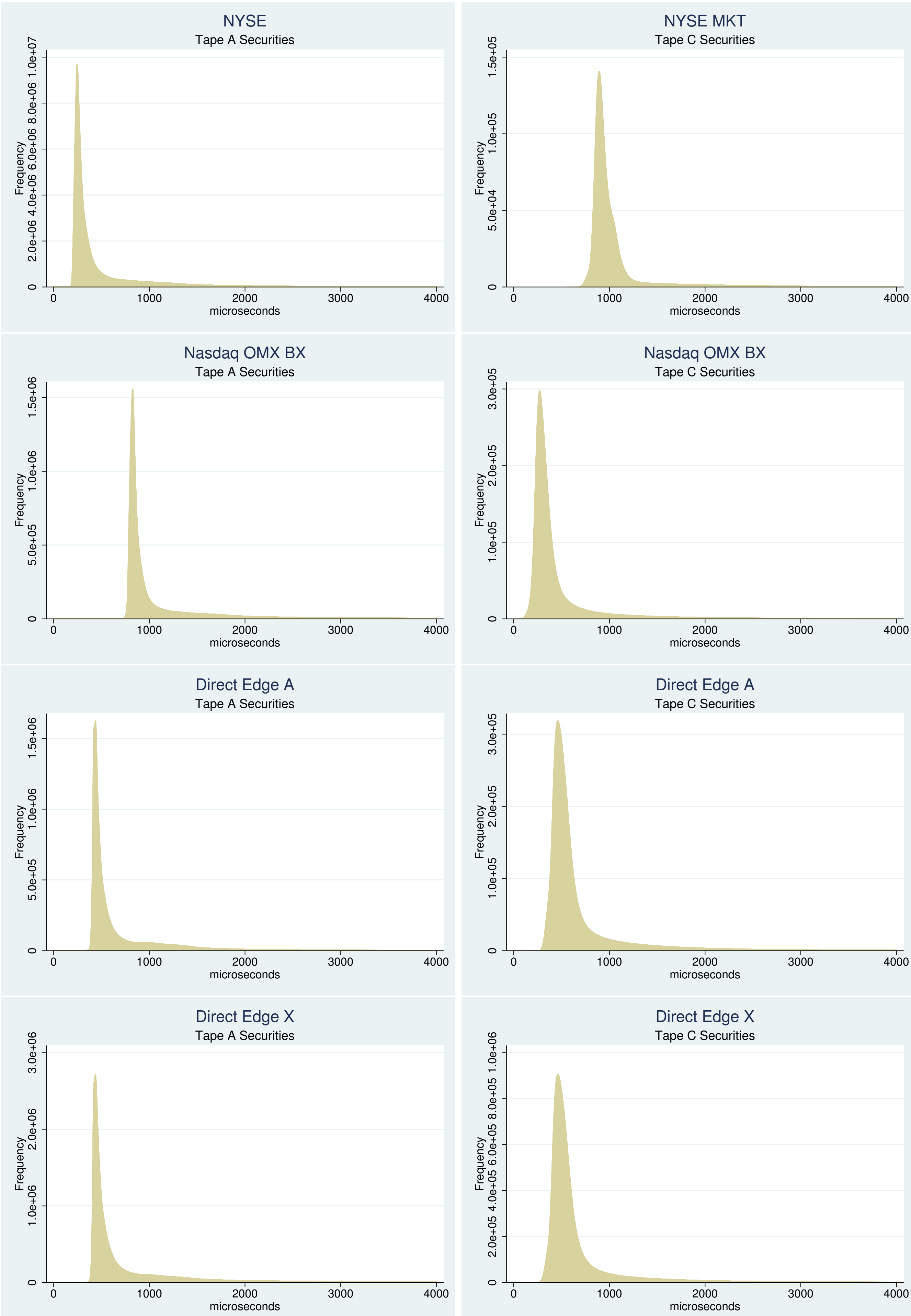
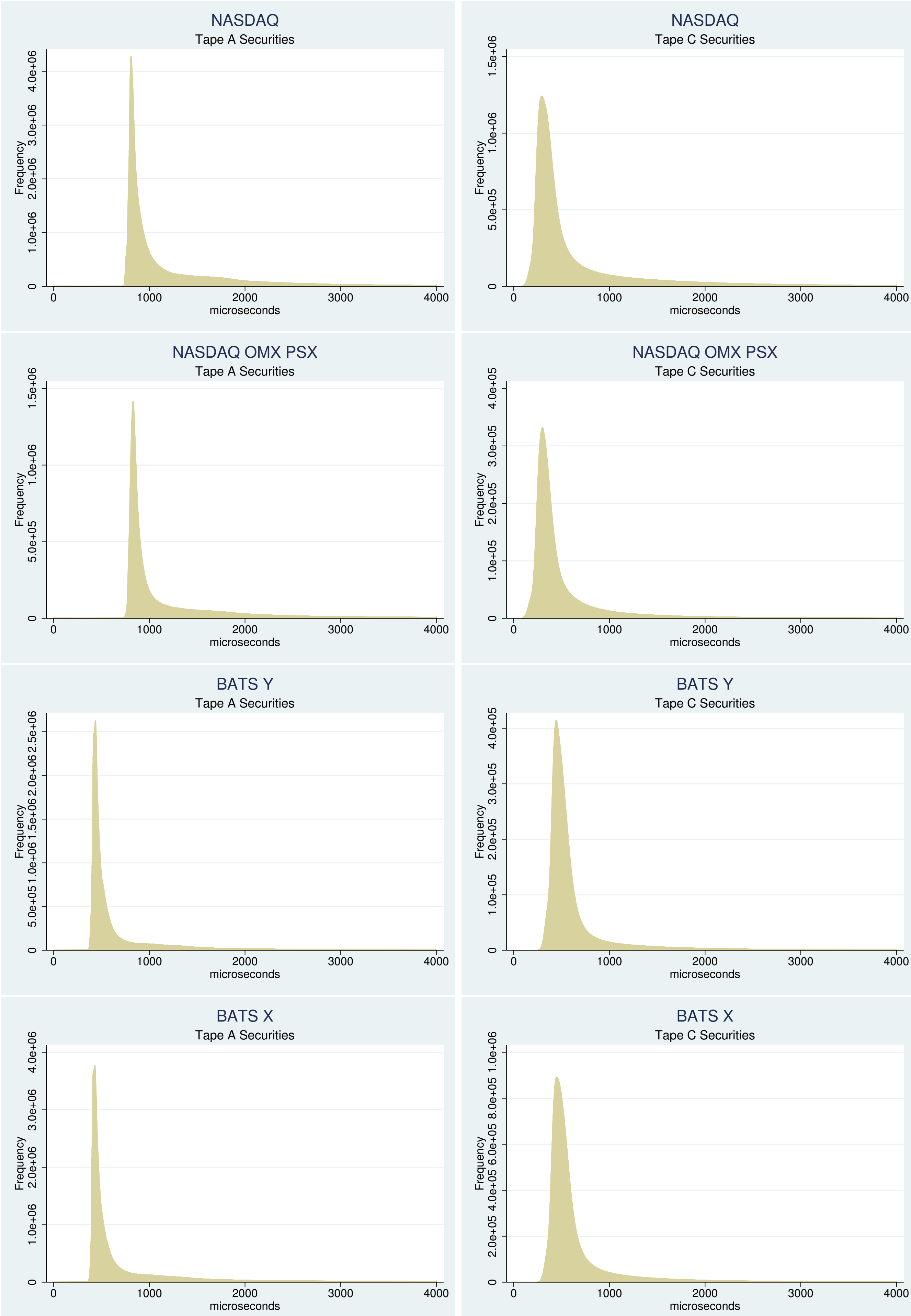
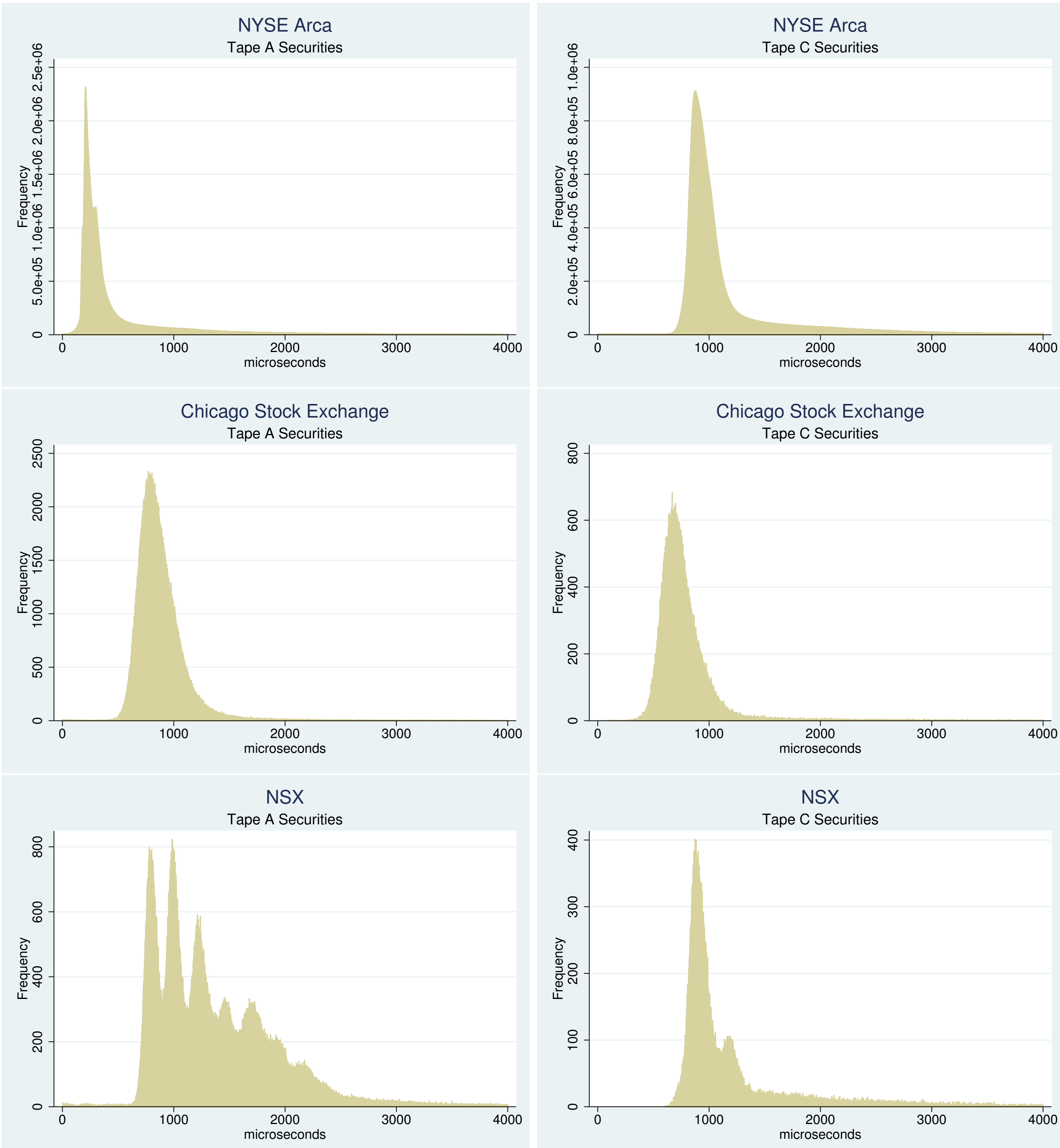


Figure A.2: Latency Distribution - Quotes Updates

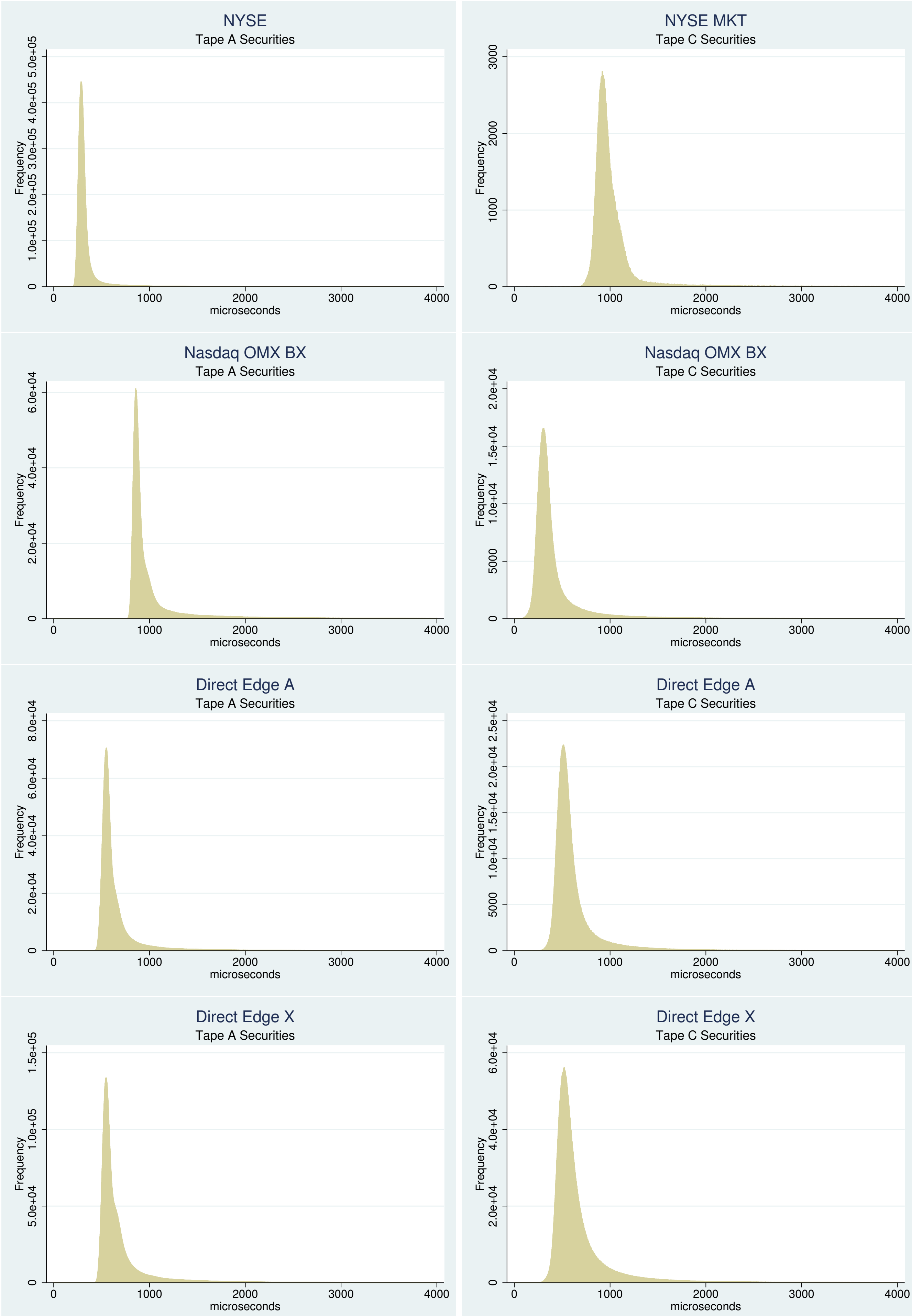


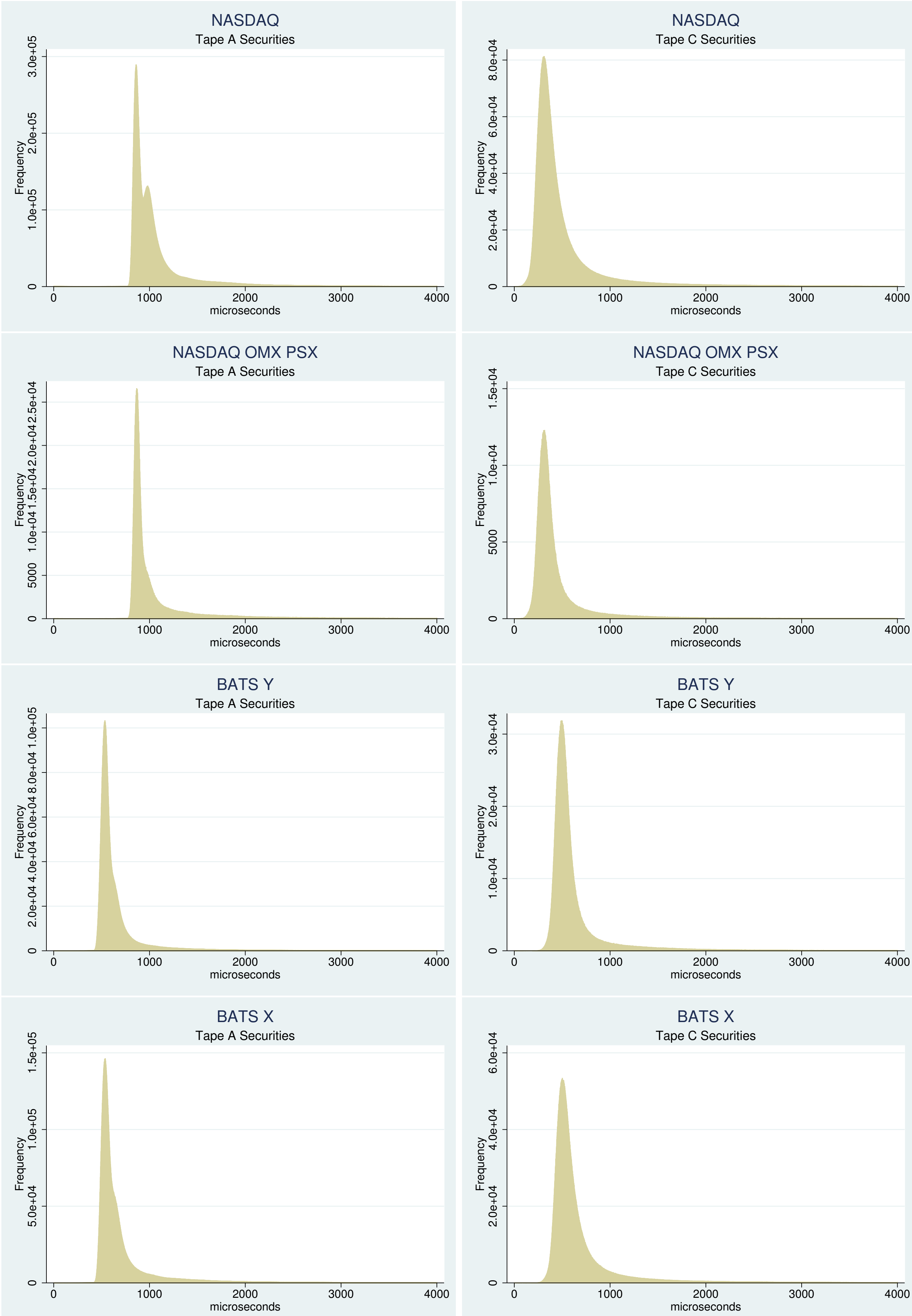

NASDAQ
Tape A Securities
Frequency
0 1.0e+06 2.0e+06 3.0e+06 4.0e+06
0 1000 2000 3000 4000
microseconds
NASDAQ
Tape C Securities
Frequency
0 5.0e+05 1.0e+06 1.5e+06
0 1000 2000 3000 4000
microseconds
NASDAQ OMX PSX
Tape A Securities
Frequency
0 5.0e+05 1.0e+06 1.5e+06
0 1000 2000 3000 4000
microseconds
NASDAQ OMX PSX
Tape C Securities
Frequency
0 1.0e+05 2.0e+05 3.0e+05 4.0e+05
0 1000 2000 3000 4000
microseconds
BATS Y
Tape A Securities
Frequency
0 5.0e+05 1.0e+06 1.5e+06 2.0e+06 2.5e+06
0 1000 2000 3000 4000
microseconds
BATS Y
Tape C Securities
Frequency
0 1.0e+05 2.0e+05 3.0e+05 4.0e+05
0 1000 2000 3000 4000
microseconds
BATS X
Tape A Securities
Frequency
0 1.0e+06 2.0e+06 3.0e+06 4.0e+06
0 1000 2000 3000 4000
microseconds
BATS X
Tape C Securities
Frequency
0 2.0e+05 4.0e+05 6.0e+05 8.0e+05 1.0e+06
0 1000 2000 3000 4000
microseconds



NYSE Arca
Tape A Securities
Frequency
0
5.0e+05
1.0e+06
1.5e+06
2.0e+06
2.5e+06
0
1000
2000
3000
4000
microseconds
NYSE Arca
Tape C Securities
Frequency
0
2.0e+05
4.0e+05
6.0e+05
8.0e+05
1.0e+06
0
1000
2000
3000
4000
microseconds
Chicago Stock Exchange
Tape A Securities
Frequency
0
500
1000
1500
2000
2500
0
1000
2000
3000
4000
microseconds
Chicago Stock Exchange
Tape C Securities
Frequency
0
200
400
600
800
0
1000
2000
3000
4000
microseconds
NSX
Tape A Securities
Frequency
0
200
400
600
800
0
1000
2000
3000
4000
microseconds
NSX
Tape C Securities
Frequency
0
100
200
300
400
0
1000
2000
3000
4000
microseconds

Figure A.3: Latency Distribution - Trades




NASDAQ
Tape A Securities
Frequency
0
1.0e+05
2.0e+05
3.0e+05
0
1000
2000
3000
4000
microseconds
NASDAQ
Tape C Securities
Frequency
0
2.0e+04
4.0e+04
6.0e+04
8.0e+04
0
1000
2000
3000
4000
microseconds
NASDAQ OMX PSX
Tape A Securities
Frequency
0
5000
1.0e+04
1.5e+04
2.0e+04
2.5e+04
0
1000
2000
3000
4000
microseconds
NASDAQ OMX PSX
Tape C Securities
Frequency
0
5000
1.0e+04
1.5e+04
0
1000
2000
3000
4000
microseconds
BATS Y
Tape A Securities
Frequency
0
2.0e+04
4.0e+04
6.0e+04
8.0e+04
1.0e+05
0
1000
2000
3000
4000
microseconds
BATS Y
Tape C Securities
Frequency
0
1.0e+04
2.0e+04
3.0e+04
0
1000
2000
3000
4000
microseconds
BATS X
Tape A Securities
Frequency
0
5.0e+04
1.0e+05
1.5e+05
0
1000
2000
3000
4000
microseconds
BATS X
Tape C Securities
Frequency
0
2.0e+04
4.0e+04
6.0e+04
0
1000
2000
3000
4000
microseconds

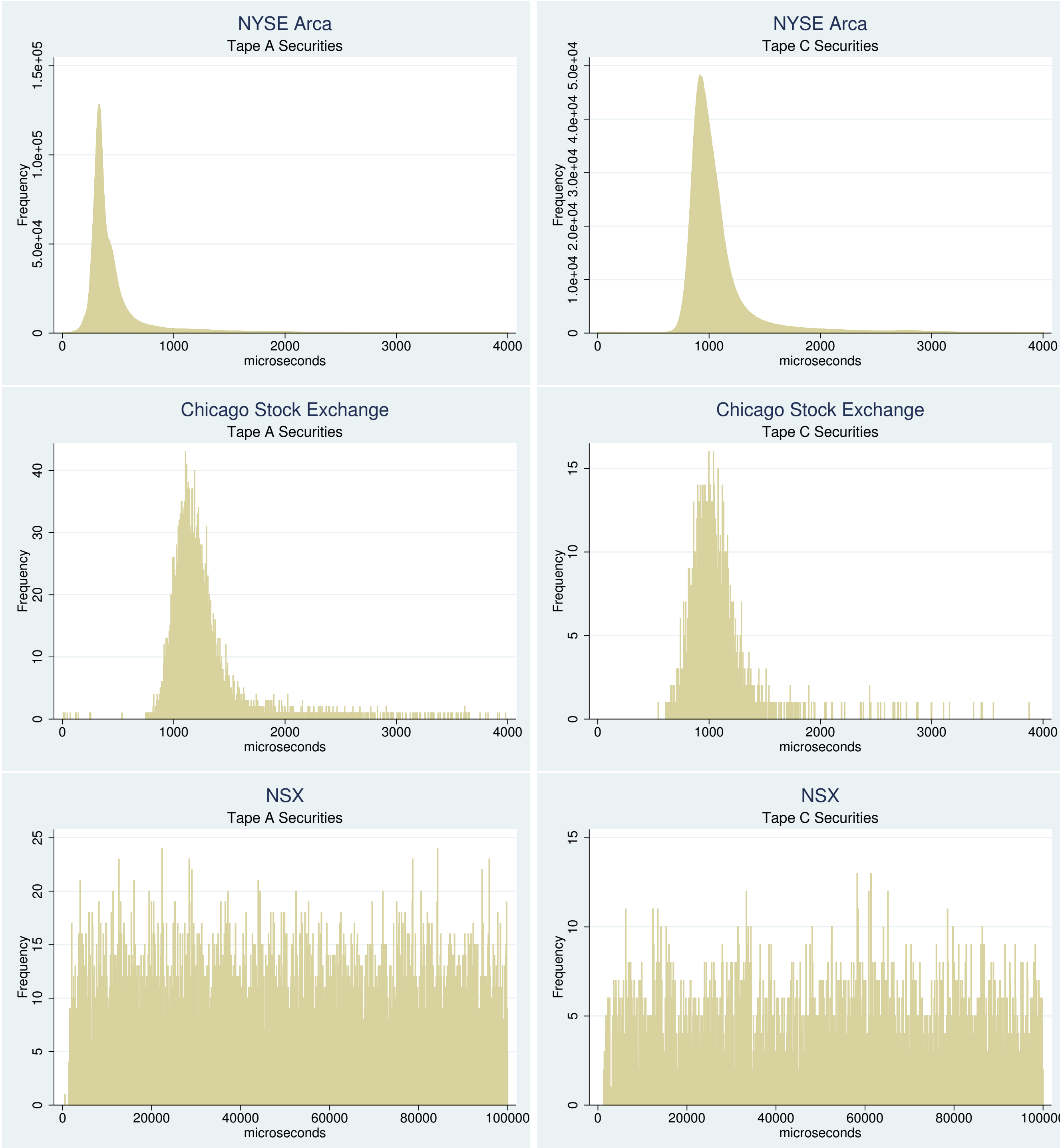
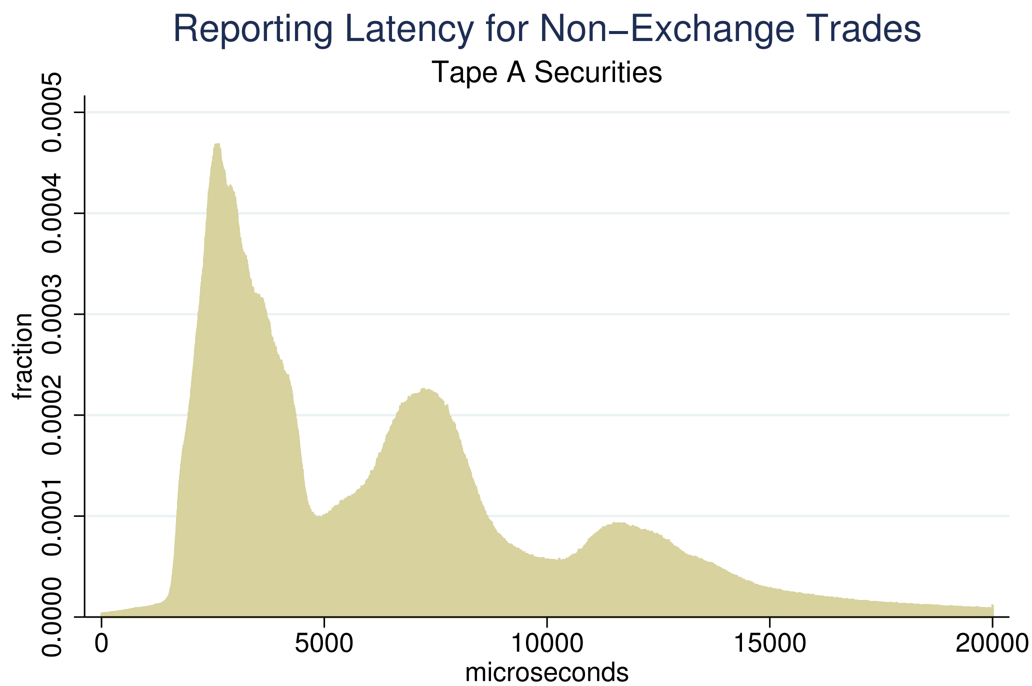
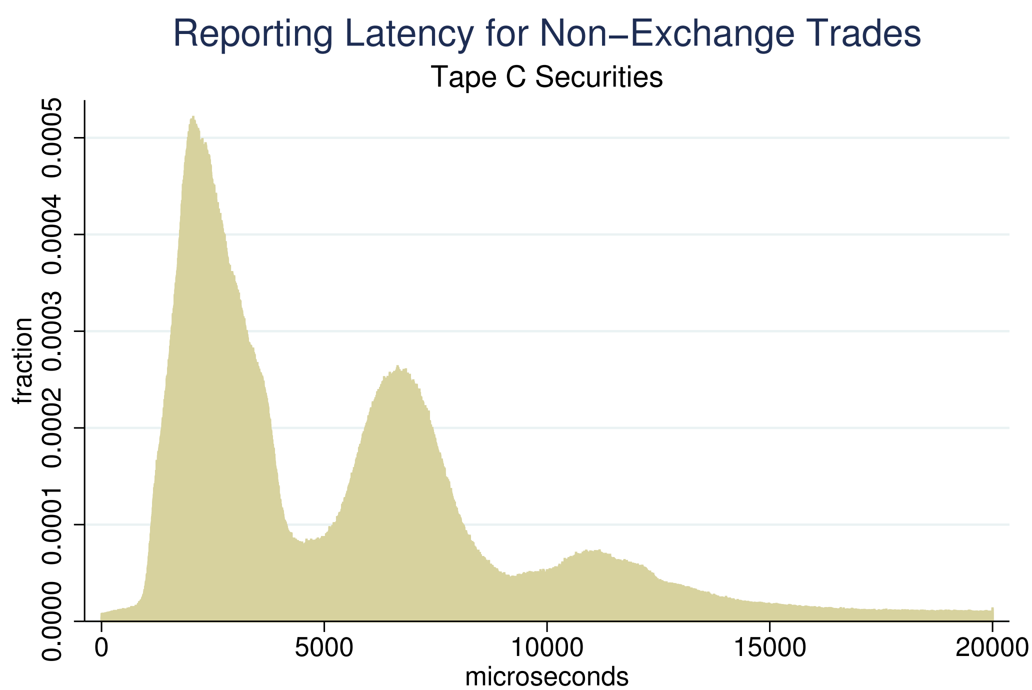

NYSE Arca
Tape A Securities
Frequency
0
5.0e+04
1.0e+05
1.5e+05
0
1000
2000
3000
4000
microseconds
NYSE Arca
Tape C Securities
Frequency
0
1.0e+04
2.0e+04
3.0e+04
4.0e+04
5.0e+04
0
1000
2000
3000
4000
microseconds
Chicago Stock Exchange
Tape A Securities
Frequency
0
10
20
30
40
0
1000
2000
3000
4000
microseconds
Chicago Stock Exchange
Tape C Securities
Frequency
0
5
10
15
0
1000
2000
3000
4000
microseconds
NSX
Tape A Securities
Frequency
0
5
10
15
20
25
0
20000
40000
60000
80000
100000
microseconds
NSX
Tape C Securities
Frequency
0
5
10
15
0
20000
40000
60000
80000
100000
microseconds

Figure A.4



Note: Roughly 20 percent of observations are above 20,000.

Figure A.5



Note: Roughly 30 percent of observations are above 20,000.