How Prevalent and Profitable are Latency Arbitrage Opportunities on U.S. Stock Exchanges?

Elaine Wah*

February 8, 2016

Abstract

In this study, I examine the prevalence of latency arbitrage opportunities that arise due to the fragmentation of trading across multiple exchanges. I analyze order and quote data from the U.S. Securities and Exchange Commission's Market Information Data Analytics System (MIDAS), which aggregates consolidated feeds and direct proprietary feeds from each U.S. stock exchange. This paper provides evidence that high-frequency traders have numerous opportunities to realize profits from latency arbitrage. These opportunities are significantly more prevalent in larger stocks and on certain exchanges. I estimate that total potential profit from latency arbitrage opportunities in S&P 500 ticker symbols was approximately $3.03 billion in 2014.

JEL classification: G10.

Keywords: high-frequency trading; latency arbitrage; market data.

*The author would like to thank Amy Edwards, Austin Gerig, Gregg Berman, Chuck Collver, Kapil Gogia, and Steve Samson for their assistance with this project and with using MIDAS data; the author is also grateful to Michael Wellman, Uday Rajan, and members of the Strategic Reasoning Group at the University of Michigan for their helpful comments and feedback. The Securities and Exchange Commission, as a matter of policy, disclaims responsibility for any private publication or statement by any of its employees or former employees. The views expressed herein are those of the author and do not necessarily reflect the views of the Commission or of the author's colleagues on the staff of the Commission. Mailing address: University of Michigan, Computer Science and Engineering, 2260 Hayward Street, Ann Arbor, MI 48109-2121. Email: ewah@umich.edu.

1. Introduction

Latency arbitrage has been used to describe a class of high-frequency trading strategies that utilize speed advantages to exploit price differences in identical (or similar) securities across multiple markets. Whether these strategies are beneficial or harmful to investors is open for debate. Before considering potential solutions to mitigate any adverse effects of such strategies, however, it is imperative to gauge the economic significance of latency arbitrage—that is, how often does it happen and how profitable is it? Quantifying the extent of such behavior has been difficult due lack of requisite data. By analyzing microsecond-level direct and consolidated feed data timestamped at a single collection point, I provide empirical evidence of the existence of latency arbitrage opportunities across 11 U.S. stock exchanges. These opportunities offer the potential for significant profit.

High-frequency trading, or HFT, has been the subject of intense scrutiny in recent years, in particular following the publication of *Flash Boys* in April 2014. HFT is generally characterized by high daily trading volume, extremely short holding periods, and liquidation to avoid significant open positions overnight (Wheatley, 2010). It is highly computerized and relies on fast access to trading platforms and market information. As such, HFT firms invest heavily in technological improvements and system upgrades to reduce their latency, the time required to access and respond to new market information.

These traders gain latency advantages through various practices that grant them access to potentially market-moving information ahead of other market participants. One such method is co-location, in which trading firms pay a premium to place their computer servers in the same data centers that house exchanges' matching engines. Trading firms may also pay for direct feeds from exchanges or invest in specialized communication lines to receive market data ahead of other traders (Goldstein et al., 2014). The TABB Group estimates that firms spent approximately $1.5 billion on technology to reduce latency in 2013 alone (Patterson, 2014). However, firms may spend millions of dollars to reduce latency only to have their systems rendered obsolete by the emergence of new technologies—perpetuating a latency arms race in which HFTs compete to access and respond to information faster than other traders. Spread Networks famously spent over $300 million in 2010 to construct a fiber optics cable for relaying market information more quickly, only to be superseded just two years later by a new microwave-based communication network (Adler, 2012).

High-frequency traders often utilize their fast access to the market to exploit short-lived arbitrage opportunities (Biais and Foucault, 2014). Exploitable price differentials may arise for a variety of reasons; in latency arbitrage, or what Lewis (2014) referred to as slow-market arbitrage, these price disparities emerge due to the fragmented nature of markets today. Market fragmentation is the phenomenon in which multiple trading venues compete for order flow. The current trading landscape is comprised of dozens of trading venues, each aggregating orders and quoting prices individually O'Hara and Ye (2011).

U.S. securities regulations have attempted to address fragmentation via Regulation NMS, which mandates cross-market communication and order routing for best execution (Blume, 2007; U.S. Securities and Exchange Commission, 2005). An entity called the Securities Information Processor (SIP) consolidates order information from exchanges to generate a public price quote, or the National Best Bid and Offer (NBBO). Submitted orders may be routed in order to avoid execution at prices worse than those available in other markets. Due to latencies in computing and updating this public price quote, the NBBO is not always up to date. Via cutting-edge technology and access to direct feeds from exchanges, HFTs can compute the best prices available before the SIP has updated the NBBO. A speed-advantaged trader who is able to see price movements in one market before prices have changed in other venues can then respond accordingly before other traders (Gaffen and Curran, 2012). As a result, a high-frequency latency arbitrageur may readily exploit disparities across fragmented markets before they are reflected in the public ticker.

The debate around the impact of latency arbitrage on investors is still ongoing. Lewis (2014) and Arnuk and Saluzzi (2012) argue that faster access to market information has opened the door to exploitative tactics by speed-advantaged traders. Baron et al. (2012)

provide evidence that HFT activities derive their profits primarily from non-HF market participants and investors. On the other hand, some posit that HFT arbitrageurs make markets more fair to retail investors by ensuring prices are properly aligned across fragmented markets, producing more accurate prices (Angel and McCabe, 2013). Others claim that the economic value of latency arbitrage is minimal and that markets "uncross" themselves, therefore such arbitrage opportunities are short-lived and self-correcting (Narang, 2014).

Given the ongoing debate, questions regarding the economic significance and potential impact of latency arbitrage are of foremost importance. However, the extent and prevalence of these exploitable price differentials in today's markets are not completely clear. Arnuk and Saluzzi (2009) estimate that latency arbitrage and other so-called "predatory" HFT strategies generate approximately \$6 to \$12 million a day, which is equivalent to \$1.5 to 3 billion in a year, but this is a rough approximation. An estimate from the TABB Group put annual aggregate profits from latency arbitrage in excess of \$21 billion in 2009 (Schneider, 2012). More recently, FINRA identified more than 395,000 transactions in Sigma X, a dark pool operated by Goldman Sachs, that executed at prices inferior to the NBBO during one week in Summer 2011 (Levine, 2014).

The most salient previous work in this vein is by Ding et al. (2014), who demonstrate that latencies in updating the NBBO allow high-frequency traders to calculate a synthetic NBBO. The authors use the NASDAQ SIP feed and direct-feed data from five exchanges to construct a synthetic NBBO. They show that the price dislocations between the synthetic NBBO and the official NBBO may be exploited by HFTs for profit. This study focuses on a type of latency arbitrage in which an HFT who knows that the NBBO is out of date submits orders to both a dark pool and an exchange in order to net a profit at the expense of the dark pool investor. Speed advantages are useful in this setting because many dark pools match orders at midpoint of the SIP NBBO. The authors analyze one day of Apple stock data to generate an upper bound of on latency arbitrage profits of approximately \$32,510 in a single stock on a single day.

Budish et al. (2015) examine arbitrage opportunities between two securities that track the S&P 500 index using millisecond-level direct feed data from the Chicago Mercantile Exchange and the New York Stock Exchange. They demonstrate that correlation between the iShares SPDR S&P 500 exchange traded fund (SPY) and the E-Mini Future (ES) breaks down at high-frequency timescales because the prices do not move simultaneously. This correlation breakdown creates technical arbitrage opportunities that the authors estimate are worth approximately $75 million per year in ES-SPY alone.

Other prior work on latency arbitrage has focused on analytical models of HFT behavior. Cohen and Szpruch (2012) analyze a model of latency arbitrage in a limit order book with a fast and a slow trader. Jarrow and Protter (2012) construct a model to show that HFT activity can lead to mispricings, creating abnormal profit opportunities for the high-frequency traders that they can then exploit.

In this paper, I focus on a specific form of latency arbitrage in which price disparities arise due to market fragmentation and delays in updating the public price quote. The type of latency arbitrage studied in this work is most closely related to that described by Wah and Wellman (2013), who employ an agent-based simulation model of two markets to demonstrate that latency arbitrage—as facilitated by order routing and fragmentation—has the potential to significantly degrade allocative efficiency.

This work is the first empirical study of the extent and prevalence of cross-market latency arbitrage opportunities across a representative sample of U.S. stock exchanges. I quantify the number, duration, and profitability of these cross-exchange latency arbitrage opportunities using exchange data from the Market Information Data Analytics System (MIDAS), a platform at the U.S. Securities and Exchange Commission (SEC) that provides access to order messages on all U.S. stock exchanges (U.S. Securities and Exchange Commission, 2013). The U.S. equities data available on MIDAS includes every trade and quote broadcast by the exchanges and SIPs (Popper and Protess, 2013). I identify cross-market arbitrage opportunities by analyzing MIDAS direct-feed data from 11 U.S. equities exchanges. My dataset includes order and quote messages between January 1, 2014 and December 31, 2014 for 495 tickers from the S&P 500 and 46 tickers from the Russell 2000. I find that latency arbitrage opportunities across exchanges arise frequently for S&P 500 stocks. These arbitrage opportunities have the potential for significant profit: I estimate that the total realizable profit for S&P 500 stocks exceeded $3.03 billion in 2014.

This paper proceeds as follows. Section 2 describes the data from MIDAS used for this study. Section 3 presents the definition of a latency arbitrage opportunity I use in my analysis. Section 4 discusses my results regarding the prevalence and profitability of these opportunities, and I conclude in Section 5.

2. MIDAS

For this study, I use order and quote data from the U.S. Securities and Exchange Commission's Market Information Data Analytics System, or MIDAS (U.S. Securities and Exchange Commission, 2013). MIDAS is a cloud-based system—accessible only from within the SEC firewall—that offers data feeds from all U.S. stock exchanges and the SIP consolidated feeds. It is developed by Tradeworx, a high-speed trading firm that also offers financial technology and market access services (Popper and Protess, 2013). MIDAS provides full depth-of-book coverage, including all order and trade messages on all U.S. stock exchanges, such as order submissions, modifications, and cancellations.[1] It does not provide information on order routing, and it does not include identifying information on market participants. As a result of this anonymization, it is also not possible to use MIDAS data to ascertain whether two orders on different exchanges were submitted by the same trader. Although some off-exchange

[1] Exchanges use two formats to report trade and order activities on their direct feeds: order-based and level-book. NYSE and NYSE MKT use the level-book format, but all 11 other exchanges in Table 1 use the order-based format. Certain metrics computed from order-based feeds, such as those based on the number of trades, may not be suited for direct comparison with those derived from level-book feeds, since level-book updates aggregate activities at specific price levels. The order-based method is more granular as it includes messages for every displayed quote or order (U.S. Securities and Exchange Commission, 2014a). The definition of a latency arbitrage opportunity used in this study is not affected by level-book versus order-based reporting, as I only analyze top-of-book prices on each exchange in addition to NBBO quotes.

Exchange	**% of market in January 2014**
NASDAQ Stock Market	26.60
New York Stock Exchange (NYSE)	16.59
NYSE Arca	15.87
BATS BZX Exchange	13.16
BATS EDGX Exchange	12.38
NASDAQ OMX BX	4.63
BATS EDGA Exchange	4.18
BATS BYX Exchange	3.11
National Stock Exchange (NSX)	0.95
NASDAQ OMX PSX	0.75
Chicago Stock Exchange (CHX)	0.68
CBOE Stock Exchange (CBSX)	0.66
NYSE MKT (Amex)	0.44

Table 1: Table of U.S. stock exchanges operational for all or part of 2014. Each exchange's percentage of the equities exchange market is computed using data from BATS Global Markets (2015b) as the percentage of mean daily trading volume in January 2014.

activities (i.e., executions on certain venues via the SIP feeds) are available on MIDAS, I focus solely on trading activity on exchanges.

MIDAS also aggregates data from the consolidated feeds. The consolidated feeds report the top-of-book quotes, or the price and total shares available for the highest buy order (BID) and the lowest sell order ($OFFER$). These quotes make up the NBBO.[2]

Each message in MIDAS is a single, atomic event on an exchange or a SIP feed that is timestamped to the microsecond by Tradeworx. Message timestamps indicate the time at which the event was recorded. Events are also timestamped by the exchanges, and MIDAS provides the difference between the Tradeworx timestamp and the exchange timestamp for each message. This is a unique advantage to MIDAS data, as messages timestamped at a single collection point eliminate the need for clock synchronization—otherwise a necessity

[2]Not all BID and $OFFER$ quotes are reported to the consolidated tape, as the consolidated system only reports round lots. MIDAS includes odd lots from the direct feeds only. Therefore, there may be discrepancies between the NBBO quote and the visible BIDs and $OFFER$s available if the total quantity available at the best price does not meet the 100-share round lot threshold (Hasbrouck, 2010). Some venues may aggregate odd lot quotes to create round lot quotations, whereas others do not use odd lot quotations in the best quotes reported to the SIP (U.S. Securities and Exchange Commission). However, my analysis does not depend on computation of a synthetic NBBO, so the irregularities in the NBBO generation should have minimal impact on my results.

when combining market data from multiple sources. Such timestamps also best capture the perspective of a high-frequency trader with direct feeds to multiple exchanges. Therefore, I restrict my analysis to the Tradeworx timestamps.

My dataset is comprised of order and quote messages from the consolidated and exchange market data feeds. I include SIP quote feeds from MIDAS: the Consolidated Tape Association (CTA) Quotes (Tapes A and B) and the Unlisted Trading Privilege (UTP) Quotes (Tape C). I also include direct-feed data from the 11 U.S. stock exchanges in operation throughout 2014 and 2015. This includes NASDAQ, NASDAQ OMX BX, NASDAX OMX PSX, New York Stock Exchange (NYSE), NYSE Arca, NYSE MKT (Amex), BATS Exchange (BZX), BATS Y-Exchange (BYX), EDGA Exchange, EDGX Exchange, and Chicago Stock Exchange (CHX). Two additional exchanges were operational for less than half of 2014 and thus are excluded from the dataset: The CBOE Stock Exchange (CBSX) ended trading operations on April 30, 2014 (CBOE Stock Exchange, 2014), and the National Stock Exchange (NSX) ceased operations on May 30, 2014 (U.S. Securities and Exchange Commission, 2014b) although it later re-launched in late December 2015 (National Stock Exchange, 2015). CBSX and NSX accounted for a combined 1.61% of the U.S. equities exchange market (Table 1), a very small fraction of the total market. As such, their exclusion from my dataset has minimal impact on overall prevalence and profitability of latency arbitrage opportunities across the other 11 exchanges.

I analyze direct feed messages from a total of 541 ticker symbols: 495 tickers from the S&P 500 and 46 tickers from the Russell 2000 index. The S&P 500 tickers selected are based on index constituents as of December 31, 2014, and I only include tickers with data for the entirety of 2014. The set of excluded S&P 500 symbols is listed in Table 2.

I include 46 tickers from the Russell 2000 index in order to examine how the prevalence of latency arbitrage opportunities differs in stocks with lower market capitalization. These stocks are based on index membership as of June 26, 2015. From the full list of index constituents, I select stocks ranked 1451 to 1500 in market capitalization (computed from

Ticker symbols				
ANTM	ES	GOOGL	KHC	NAVI
TGNA	WBA	WRK	ZBH	

Table 2: S&P 500 tickers excluded for lacking direct feed data for the entirety of 2014. The S&P 500 dataset is comprised of order and quote messages for 495 S&P 500 ticker symbols (based on membership as of December 31, 2014) with data for all of 2014.

Ticker symbols							
MRCY	MSFG	WINA	GHM	TITN	SNBC	TTPH	RGLS
ATRS	EXAC	HNH	USCR	NLS	CRMT	ISRL	GBLI
RMAX	AAOI	ETM	PTX	OLP	ISLE	MSEX	IMMU
BCOV	OKSB	FC	HSII	CSBK	BBSI	AFOP	UVSP
PPHM	CENTA	LNDC	FPRX	CTRN	MCS	WSR	NILE
VCRA	TREE	SGI	AE	XOXO	RTEC		

Table 3: Russell 2000 tickers included in the dataset. The Russell 2000 dataset is comprised of order and quote messages for 46 ticker symbols (based on membership in Russell 2000 as of June 26, 2015) with data for all of 2014.

closing prices and number of shares outstanding on June 30, 2014, obtained from the Center for Research in Security Prices). This is to avoid any potential volatility in lower-cap stocks before the Russell index reconstitution. As with the S&P 500 tickers, I only include tickers with data for the entirety of 2014. The list of Russell 2000 stocks included is given in Table 3.

My dataset includes consolidated and exchange data feed messages for the 251 trading days between January 1, 2014 and December 31, 2014 (inclusive), and is restricted to messages received during trading hours (i.e., 9:30 am to 4:00 pm Eastern Time). I exclude market holidays and for early closures I restrict analysis to messages received between 9:30 am and 1:00 pm Eastern Time.

I also exclude October 21, 2014 from my analysis. On this date NYSE MKT experienced an issue processing market quotes in 150 NASDAQ-listed symbols, 41 of which are in my dataset. The problem occurred on one of their matching engines and affected both order processing and trade execution. Trading was halted in the affected stocks, and only resumed in those symbols on the following day (NYSE, 2015). Market disruptions occurred on a number of other dates on various exchanges, but the disruption on October 21, 2014 directly affected the largest number of tickers included in the dataset. Problems with the consolidated feeds also occurred in 2014—for instance, on October 30, 2014, there was an issue reporting quotes and trades to one of the SIP feeds (Nasdaq, 2015; Chicago Stock Exchange, 2015; BATS Global Markets, 2015a)—but the exact durations of such disruptions, as well as the specific tickers affected, are unclear based on publicly available data.

3. Quantifying latency arbitrage opportunities

In this paper, I focus on latency arbitrage opportunities that arise due to market fragmentation across multiple exchanges and latencies in updating the NBBO. This form of latency arbitrage is most similar to the HFT strategy modeled by Wah and Wellman (2013), in which a latency arbitrageur exploits price disparities across fragmented markets caused by order routing based on outdated NBBO quotes.

3.1. Definition of a latency arbitrage opportunity

I define a latency arbitrage opportunity as an instance in which the highest-priced buy order (BID) on one exchange crosses the lowest-priced sell order ($OFFER$) on another exchange, and neither price is worse than the the best buy and sell prices as specified by the NBBO.

Definition 1. *Given exchanges A and B, a latency arbitrage opportunity* $OPP_{A,B}$ *exists if:*

(1) The two markets are crossed, i.e., if $BID_A > OFFER_B$.

(2) Neither side of the arbitrage opportunity trades through the NBBO, i.e., neither the BID_A *price nor the* $OFFER_B$ *price is worse than the best prices available elsewhere according the SIP:*

$$NBB \leq BID_A$$

$$NBO \geq OFFER_B,$$

where NBB is the national best buy-order price and NBO is the best sell-order price, across all exchanges.

(3) The arbitrage opportunity begins and ends on different timestamps.

Each arbitrage opportunity $OPP_{A,B}$ is uniquely defined by the two exchanges involved and by the BID_A and $OFFER_B$ prices on these exchanges. Note it should not be possible for two latency arbitrage opportunities $OPP_{A,B}$ and $OPP_{B,A}$ to exist simultaneously across the same pair of exchanges, as this would entail a crossed market in one of the two exchanges. I base my definition of a latency arbitrage opportunity on execution relative to the NBBO quote because the SIP feeds are used to assess compliance—brokers and exchanges are permitted to use SIP quotations, direct feeds, or some combination to comply with Regulation NMS (Stone, 2014).

Figure 1 illustrates an example of a latency arbitrage opportunity across exchanges A and B. Price disparities may emerge across a pair of exchanges as a result of latencies in updating the NBBO. In this example, the NBBO is out of date, and an arbitrage opportunity exists between the two exchanges. Executions at the BID on exchange A or the $OFFER$ on exchange B are not trade-throughs[3] according to the SIP NBBO. An infinitely fast latency arbitrageur with access to direct feeds to exchanges A and B can readily see that the markets are crossed. If the arbitrageur were to submit a market order to sell in exchange A and one to buy in exchange B, it would realize a profit of \$0.01 per share from this arbitrage opportunity.

The size δ of arbitrage opportunity $OPP_{A,B}$ is the spread $BID_A - OFFER_B$. Quantity q is the maximum number of shares available at both the BID and $OFFER$ prices. Duration $T > 0$ is measured from the time the price disparity is first detected to when the arbitrage opportunity ends. A given arbitrage opportunity continues as long as the conditions in Definition 1 hold and the prices at which the markets are crossed remain the same. In

[3] While it is not possible for a cross-exchange arbitrage opportunity to arise that would also trade through the NBBO, trade-through rates were estimated to be 0.11% for NASDAQ stocks and 0.13% for NYSE stocks in February 2014 (U.S. Securities and Exchange Commission, 2015), so this scenario is unlikely.

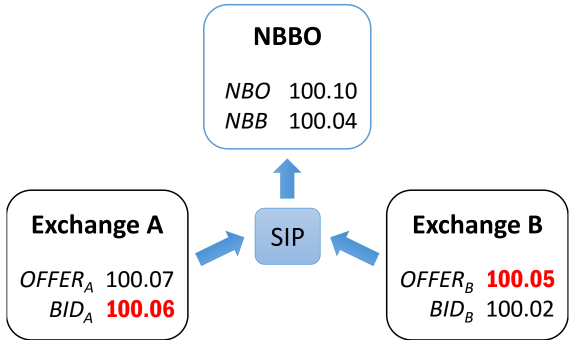


Fig. 1. Example of a latency arbitrage opportunity between two exchanges A and B. The red, bolded prices highlight a crossed market: $BID_A > OFFER_B$. Execution at BID_A or $OFFER_B$ would not trade through the NBBO quote, as both prices match or improve upon the global best prices, NBB and NBO. If $BID_A = 100.03$, however, then $NBB > BID_A$ and a trade at BID_A would be an execution at a price worse (i.e., lower) than NBB, the best buy-order price available across all exchanges. An infinitely fast arbitrageur would net a profit of $0.01 per share from this arbitrage opportunity.

this analysis, a latency arbitrage opportunity ends for one of three possible reasons: (1) the two markets are no longer crossed due to a trade or because an order is canceled, (2) the prices forming the crossed market change, creating a different arbitrage opportunity, or (3) the NBBO changes and the conditions in the second part of Definition 1 are violated. I exclude from my analysis any arbitrage opportunities that emerge and disappear on the same timestamp, although a sub-microsecond-level HFT would certainly be able to exploit such opportunities. The quantities available at BID_A or $OFFER_B$ may change, while prices remain the same, but I do not treat such instances as a different arbitrage opportunity.

3.2. Profitability of latency arbitrage opportunities

To measure the profitability of latency arbitrage, I assume the presence of a single infinitely fast latency arbitrageur (LA). When the LA detects a latency arbitrage opportunity, its strategy is to submit market orders to the exchanges involved in the cross-market arbitrage opportunity. In measuring potential profit, I track two metrics: potential profit per opportunity (π) and realizable profit. Both metrics assume zero transaction costs and no execution risk. Given q_A shares available at price BID_A and q_B shares available at price

$OFFER_B$, the per-opportunity profit π of $OPP_{A,B}$ is:

$$\min\{q_A, q_B\} \cdot (BID_A - OFFER_B).$$

Realizable profit is the trading gains the hypothetical LA could potentially achieve under the following assumptions: it responds immediately to any identified arbitrage opportunity, it selects the most profitable opportunity if multiple arise simultaneously, and its trading activities do not affect other orders currently in the order book.

I estimate the potential profit of an arbitrage opportunity based on the initial quantity at the time of emergence. I ignore changes in the quantities available at the crossed quotes. Since quantity increases indicate there are more shares available in the arbitrage opportunity, their exclusion may result in a slightly more conservative profitability estimate. A reduction in the number of shares available very close to or after the emergence of the arbitrage opportunity could lead to execution at a worse price, if the arbitrageur does not respond quickly enough. Nevertheless, preliminary analysis suggests that quantity changes in arbitrage opportunities are somewhat rare, and thus of negligible impact.

If latency arbitrage opportunities arise across multiple pairs of exchanges, the profit calculations depend on whether the opportunities are *independent*—that is, whether the BID and $OFFER$ prices involved in the crossed markets are distinct—and whether the opportunities arise on the same timestamp. Figure 2 illustrates a scenario with two independent arbitrage opportunities, one across exchanges A and B and one across exchanges A and C. Realizable profit in this setting includes the per-opportunity profits from both pairs. If the second opportunity were to arise with the same timestamp as the first, I assume the LA would execute only the most profitable arbitrage opportunity available on that timestamp. Given the microsecond-level granularity of order messages, these instances are uncommon.

Two or more latency arbitrage opportunities may involve the same quote on a given exchange. For these *dependent* opportunities, the calculation of realizable profit includes the profit from only one of the opportunities. If multiple dependent opportunities start on

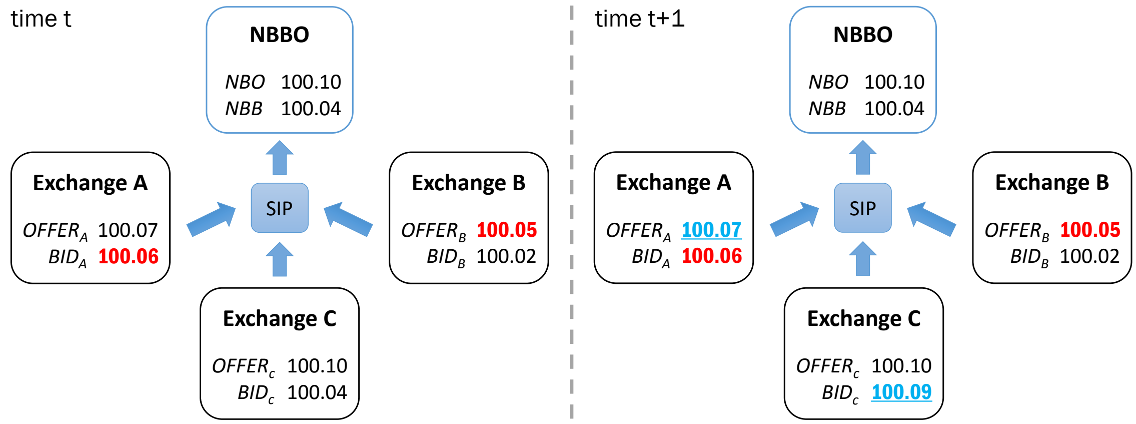


Fig. 2. Example of how per-opportunity profit and realizable profit are computed given latency arbitrage opportunities that arise across multiple exchanges. Here there are three exchanges A, B, and C. On the left at timestamp t is the latency arbitrage opportunity $OPP_{A,B}$ between two exchanges A and B described in Figure 1. The red, bolded prices highlight a crossed market between exchanges A and B: $BID_A > OFFER_B$. The potential profit from $OPP_{A,B}$ is \$0.01 per share. On the right, at time $t+1$, the BID on exchange C increases to \$100.09. There is now a second latency arbitrage opportunity $OPP_{C,A}$ across exchanges C and A. The blue, underlined prices highlight this opportunity, in which $BID_C > OFFER_A$. Execution at BID_C or $OFFER_A$ would not trade through the NBBO quote, as both prices match or improve upon NBB and NBO. The potential profit available to the infinitely fast latency arbitrageur that capitalizes on opportunity $OPP_{B,C}$ is \$0.02 per share. In computing realizable profit, I include both the profit from $OPP_{A,B}$ and from $OPP_{C,A}$, as execution of either opportunity does not affect the other.

the same timestamp, I include the arbitrage opportunity with the highest potential profit in the computation of realizable profit. Otherwise, I assume the LA executes the first of the dependent opportunities to arise. Figure 3 illustrates such a scenario. In this example, the BID on exchange A is involved in two arbitrage opportunities. Because the arbitrage across exchanges A and B arises before the arbitrage opportunity across exchanges A and C, the realizable profit metric only includes the profit from the earlier arbitrage opportunity.

For both independent and dependent arbitrage opportunities, there may be additional shares at the BID or $OFFER$ prices involved in the arbitrage. For instance, if $q_A > q_B$, upon hypothetical execution of $OPP_{A,B}$, $q_A - q_B$ shares remain at price BID_A on exchange A. While it may be possible to extract further profit in these cases, my profit estimate does not take into account such scenarios.

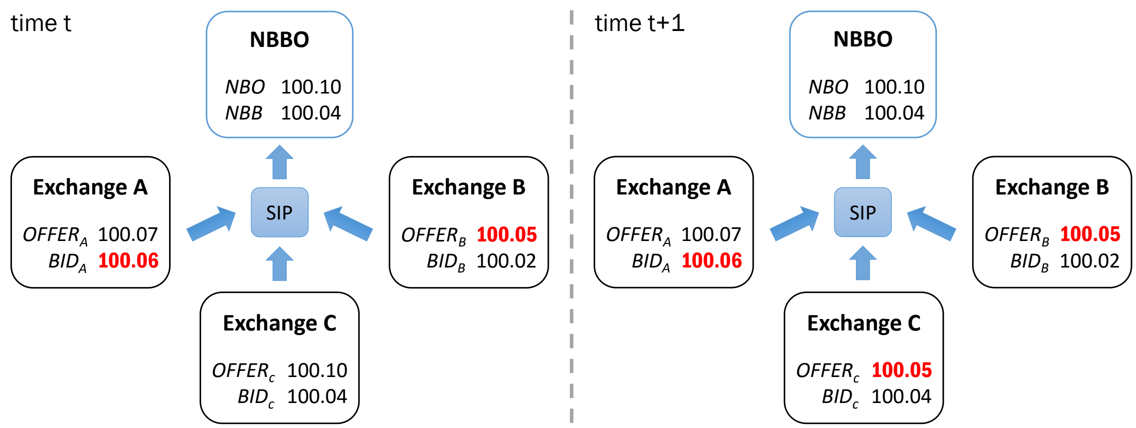
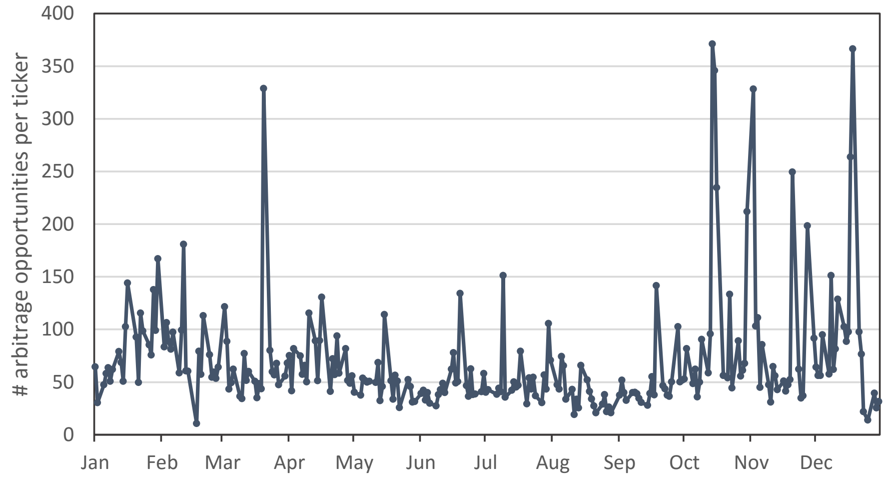


Fig. 3. Example of how per-opportunity profit and realizable profit are computed when the same quote on one exchange is involved in multiple latency arbitrage opportunities. The status at time t of the three exchanges A, B, and C is the same as in Figure 2. In this example, however, at time $t+1$ the $OFFER$ on exchange C decreases to \$100.05. There is now an arbitrage opportunity $OPP_{A,C}$ because $BID_A > OFFER_C$. Profit for $OPP_{A,C}$ is \$0.01 per share. The BID on exchange A is simultaneously involved in two arbitrage opportunities. In this scenario, realizable profit includes only the profit from the earlier arbitrage opportunity, $OPP_{A,B}$. This assumes that executing $OPP_{A,B}$ will effectively end $OPP_{A,C}$. It is possible that additional shares at BID_A are left after the trades involved in executing $OPP_{A,B}$, and further realizable profit is attainable, but my analysis does not track such scenarios.

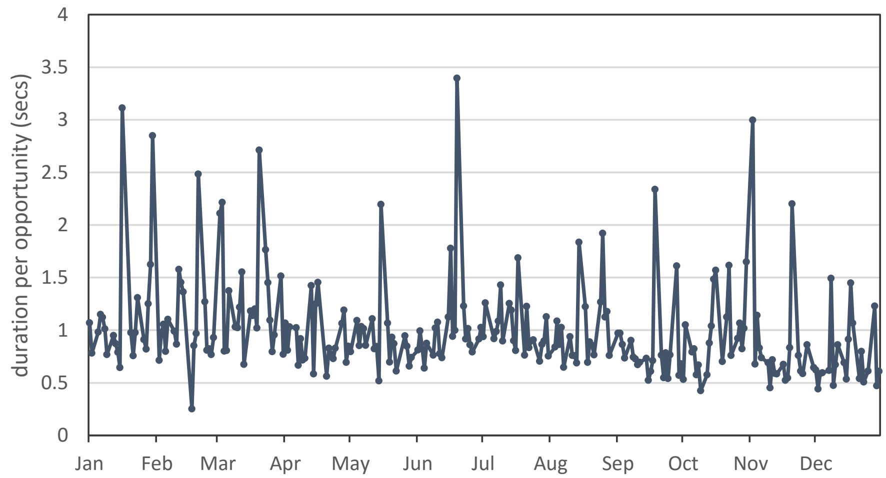
4. Results

In this section, I present my data as is.[4] Overall, I find that latency arbitrage opportunities, as defined in the previous section, are plentiful (Section 4.1). Total realizable profit from such opportunities, across the 495 S&P 500 tickers in my dataset, exceeds \$3.03 billion (Section 4.2). I find that latency arbitrage opportunities are most prevalent on NYSE and exchanges with greater market share (Section 4.3). Finally, I explore the characteristics of these opportunities in smaller stocks in the Russell 2000 index (Section 4.4), finding that latency arbitrage opportunities are less prevalent and less profitable than in larger stocks.

[4]Further analysis is not feasible, as the dataset is not accessible from outside the SEC.



(a) Mean daily number of arbitrage opportunities per ticker



(b) Mean duration per arbitrage opportunity over all tickers

Fig. 4. Latency arbitrage opportunities in the 495 S&P 500 tickers in 2014, with metrics averaged for each of the 251 trading days over the 11 exchanges in the dataset.

4.1. Latency arbitrage opportunities are plentiful

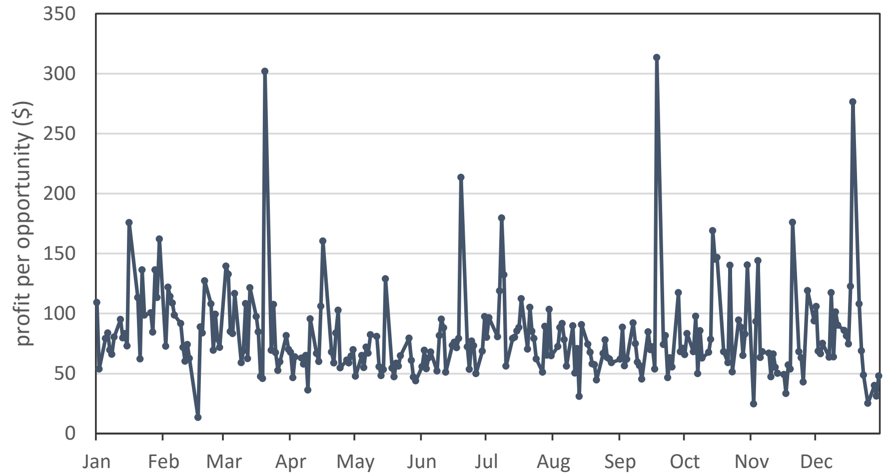
Figure 4 shows the mean daily number of latency arbitrage opportunities per ticker and the mean duration of an arbitrage opportunity, averaged for all exchanges. The number of arbitrage opportunities (Figure 4(a)) exhibits day-to-day fluctuations but otherwise remains fairly stationary for the time period in the dataset. There are approximately 69 latency arbitrage opportunities per ticker per day. The median duration, computed as the median of the daily mean per-opportunity durations (Figure 4(b)) is 0.87 seconds. The mean daily duration metric is highly sensitive to extreme values, and can be skewed upward by anomalous durations in even just a single ticker; therefore, the duration metric reported here most likely overestimates the median across all durations.

There are a number of days that exhibit an extremely high number of latency arbitrage opportunities (i.e., exceeding three standard deviations from the overall mean): March 21, October 15–17, November 3, November 21, and December 18–19. Some of these spikes may be caused by problems in an exchange data feed, which can lead to inaccuracies when quantifying latency arbitrage via direct-feed messages. The Chicago Stock Exchange reported a trading halt on November 3 due to "market data issues" (Chicago Stock Exchange, 2015). Other irregularities may be due to market disruptions or overall market conditions. For example, the U.S. Treasury market experienced a very high level of volatility within a 12-minute window on October 15 (U.S. Department of the Treasury et al., 2015), which may have affected prices in the equity markets.

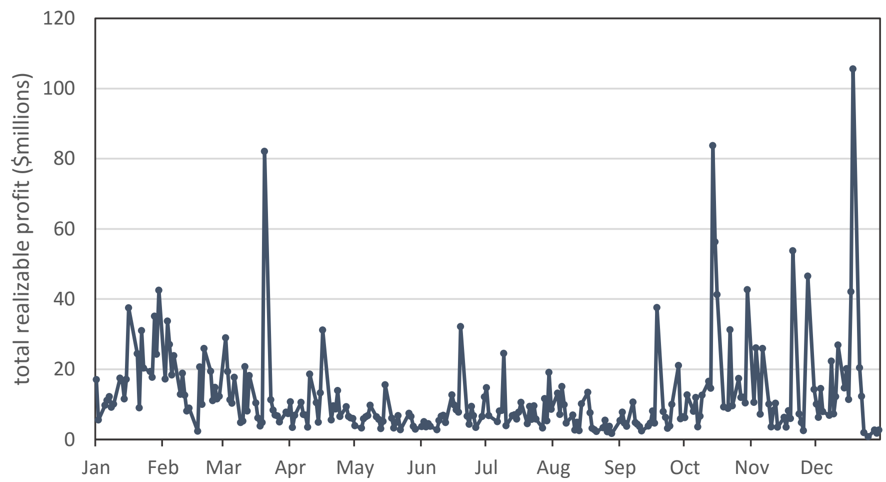
In analyzing the prevalence of latency arbitrage opportunities over time, it is possible the presence of CBSX and NSX in early 2014 may have had some effect on characteristics of latency arbitrage opportunities across the 11 exchanges in my dataset. Although I exclude direct feed data from CBSX and NSX, the presence of two additional venues could potentially alter order routing and trading activity. Combined, however, CBSX and NSX represent less than 2% of typical daily trading volume. As such, their impact on the prevalence of latency arbitrage opportunities is unlikely to have been significant. An analysis of a subset of tickers on CBSX and NSX in the period prior to their closures suggests that there are fewer latency arbitrage opportunities on these two exchanges than on exchanges with greater market share (e.g. NASDAQ and NYSE), so this assumption appears reasonable.

4.2. Latency arbitrage opportunities have high potential profit

Figure 5 shows daily profit measures—average per arbitrage opportunity as well as total realizable profit per ticker—in 2014 for the 11 exchanges in the dataset. The mean profit per arbitrage opportunity is \$81, and mean daily total realizable profit per ticker is \$24,446.



(a) Mean profit per arbitrage opportunity over all tickers



(b) Total realizable profit over all tickers

Fig. 5. Profit measures of latency arbitrage opportunities aggregated over the 495 S&P 500 tickers in 2014. Total realizable profit is the sum of realizable profit over all tickers.

There are spikes in mean profit per opportunity on March 21, October 15, and December 19, which correspond to the dates discussed in the previous section. Summing up the total potential profit for all S&P 500 stocks over the 251 trading days in the dataset, I obtain a total realizable profit of approximately $3.03 billion. This estimate is in line with previous work on the profitability of HFT strategies (Ding et al., 2014; Budish et al., 2015; Brogaard, 2010).

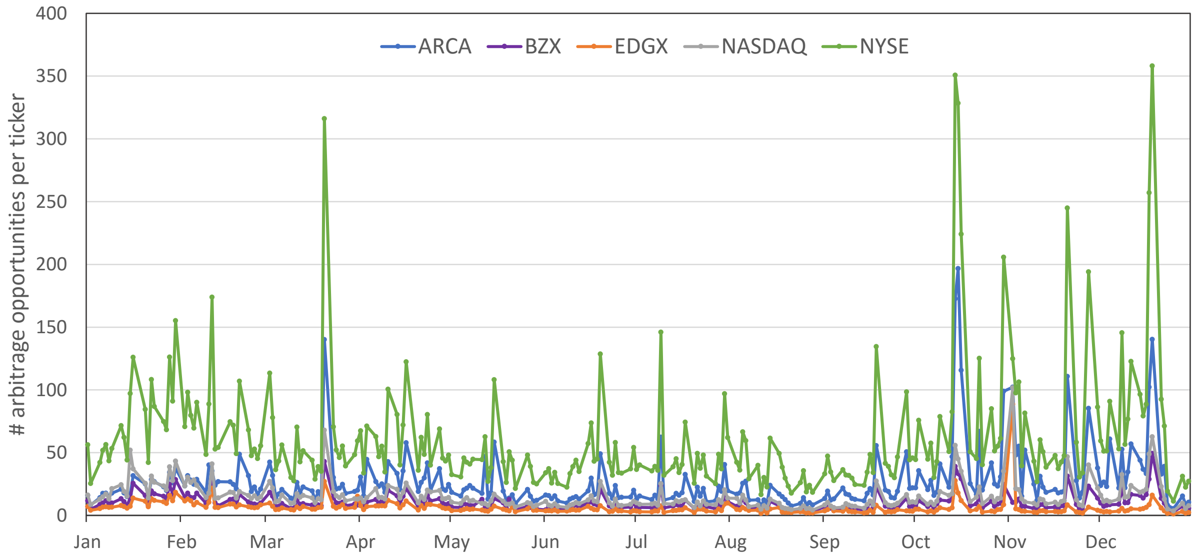
4.3. Latency arbitrage opportunities are more prevalent on certain exchanges

Figures 6 and 7 show latency arbitrage opportunities plotted by exchange. Since these are pairwise cross-exchange opportunities, each arbitrage opportunity is counted twice in these plots—once each by each market involved in the arbitrage. I observe the highest number of latency arbitrage opportunities on NYSE (Figure 6(a)), despite NASDAQ having larger market share (Table 1). This may be due to latency differences on the two exchanges. NYSE and ARCA have higher round-trip latencies (IEX Group, 2015). This stems in part from the location of their data centers in Mahwah, NJ. Higher latencies could cause the NBBO to be out of date more often, and under these circumstances latency arbitrage opportunities are more likely to arise (Wah and Wellman, 2013). The mean duration per arbitrage opportunity is also the highest on NYSE and ARCA, in line with the longer round-trip latencies. These reported latencies are from 2015, however, and the latencies in the period covered by the dataset may differ.

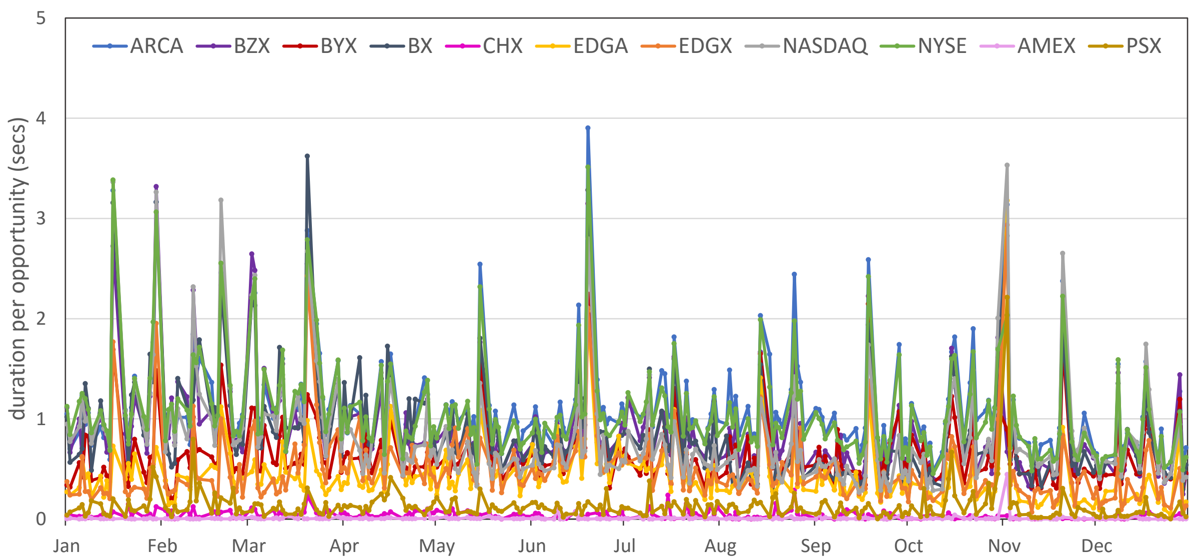
Profitability of arbitrage opportunities on each exchange corresponds closely to the number of opportunities present. Profit per arbitrage opportunity is highest on ARCA, followed closely by NYSE. Total realizable profit, summed over all 495 S&P 500 tickers, is the highest on NYSE. Unsurprisingly, most of the latency arbitrage opportunities can be found on the three exchanges responsible for over 59% of trading volume on U.S. equities exchanges. Table 4 gives a summary of prevalence and profitability metrics by exchange.

4.4. Latency arbitrage opportunities are less prevalent in smaller stocks

In this section I examine how latency arbitrage opportunities differ for smaller securities. Figure 8 shows the mean duration and number of arbitrage opportunities of the 46 stocks from Russell 2000, and Figure 9 gives the profit metrics. In contrast to latency arbitrage opportunities in the S&P 500 stocks (Figures 4 and 5), the prevalence and profitability of latency arbitrage in the Russell 2000 tickers is significantly lower. The mean number of arbitrage opportunities per ticker is 2.1, and the median of the daily mean durations in

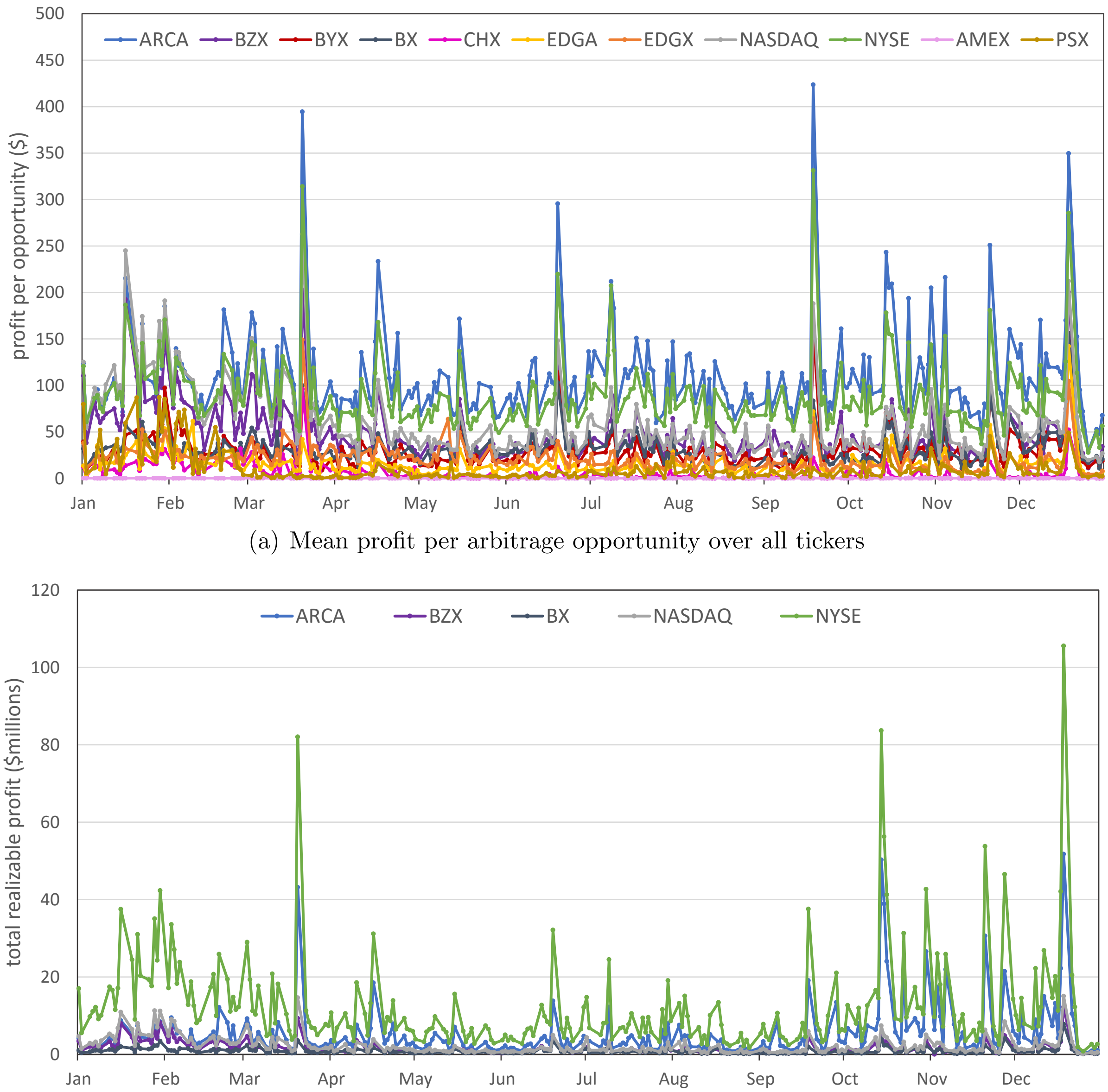


(a) Mean number of arbitrage opportunities per ticker



(b) Mean duration per arbitrage opportunity over all tickers

Fig. 6. Number and duration of latency arbitrage opportunities in the 495 S&P 500 tickers in 2014. In Figure 6(a), only the five exchanges with the highest mean number of arbitrage opportunities per ticker (averaged over 251 trading days) are shown.



(a) Mean profit per arbitrage opportunity over all tickers

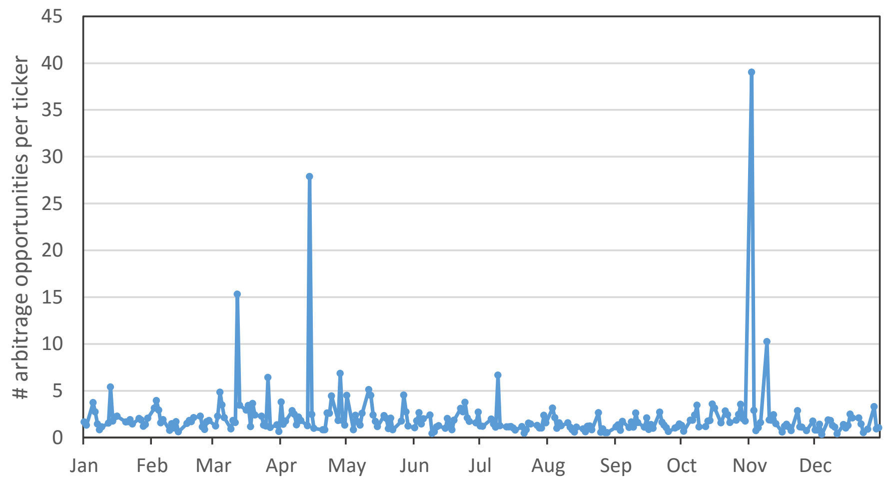
(b) Total realizable profit over all tickers

Fig. 7. Latency arbitrage opportunity profit measures for the 495 S&P 500 tickers in 2014. Figure 7(b) shows only the five exchanges with the highest mean total realizable profit per ticker, computed by averaging over the 251 trading days in the dataset.

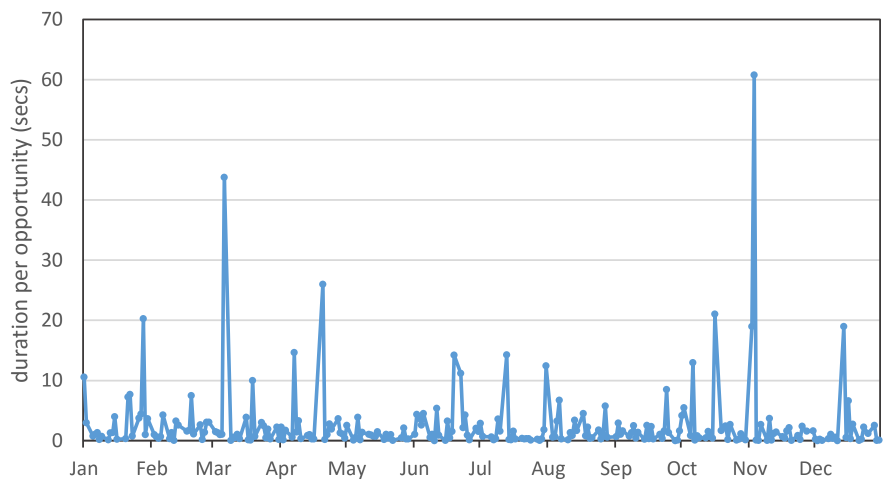
Figure 8(b) is 1.01 seconds. The average profit per opportunity is $0.75, and the mean realizable profit per ticker over all 251 trading days is $23.99. The relative results on an exchange-by-exchange basis (Table 5) are similar to those observed for the S&P 500 tickers, with the most latency arbitrage opportunities found in ARCA, NASDAQ, and BZX.

There is a significant spike in the number of latency arbitrage opportunities on March 13, April 15, and November 4, 2014. There was a spike on November 3 in the S&P 500 tickers; as discussed in Section 4.1, CHX reported a trading halt on November 3 for Tape B issues traded in Chicago (Chicago Stock Exchange, 2015), which may have also affected trading the following day. Based on the publicly available archival system alerts of the 11 exchanges in the dataset, it does not appear there was a market disruption on either March 13 or April 15 that would explain the significant increase in number of latency arbitrage opportunities. However, on both these dates, there is an abnormally high number of arbitrage opportunities in the same ticker, AE, with 628 opportunities on March 13 and over 1,000 on April 15; both of these values are well over three standard deviations from the mean. Excluding AE on those dates gives an average number of opportunities per ticker more in line with the rest of the time series in Figure 8(a). Mean durations are high, due in part to the small sample size and the inclusion of all tickers in the averaging, which renders this metric very sensitive to extreme values.

The Russell 2000 dataset includes fewer stocks than the S&P 500 dataset, and the smaller sample size means these results are more sensitive to idiosyncratic behavior in one or two stocks. However, these results strongly suggest that the larger and more actively traded stocks, such as those in the S&P 500, have significantly greater potential for latency arbitrage profits than stocks with smaller market capitalization.

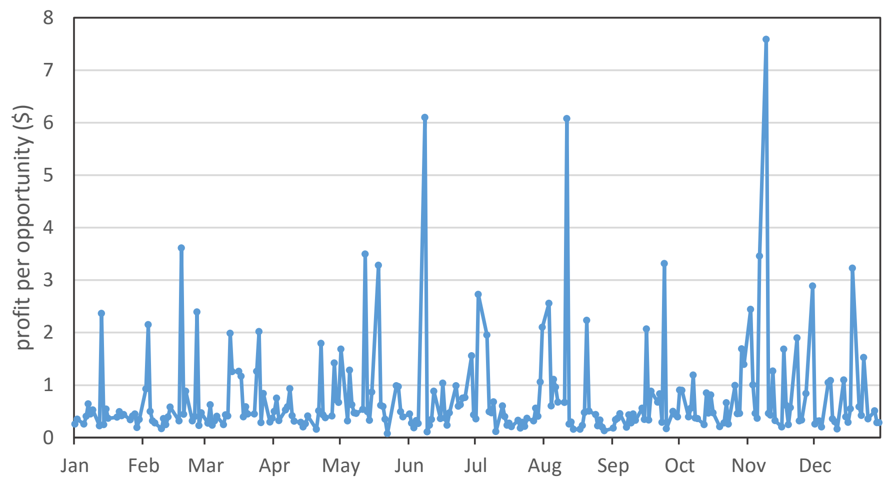


(a) Mean number of arbitrage opportunities per ticker

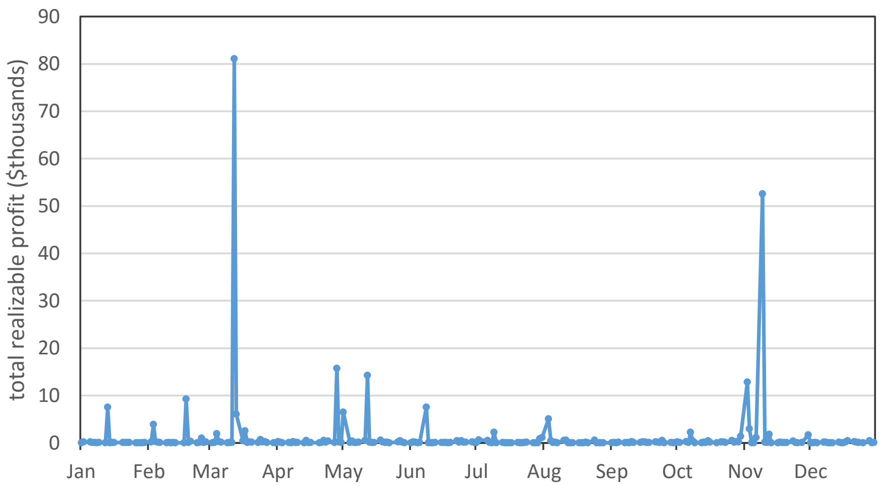


(b) Mean duration per arbitrage opportunity over all tickers

Fig. 8. Latency arbitrage opportunities in 2014 in the 46 Russell 2000 tickers in 2014, with metrics averaged over all exchanges.



(a) Mean profit per arbitrage opportunity over all tickers



(b) Total realizable profit over all tickers

Fig. 9. Profit measures for latency arbitrage opportunities in 2014 in the 46 Russell 2000 tickers, with metrics averaged over all exchanges. Total realizable profit is the sum of realizable profit over all tickers.

	BZX	BYX	CHX	EDGA	EDGX	BX	NASD	PSX	ARCA	AMEX	NYSE
N	10	6	1	4	6	6	15	2	27	1	62
T	0.75	0.55	0.03	0.37	0.40	0.74	0.68	0.10	0.98	0.003	0.94
π	48.94	29.60	5.96	15.33	21.34	31.34	59.64	10.62	108.26	0.05	88.68
π_{total}	3,061	1,697	190	883	1,220	1,761	4,542	734	10,624	4	24,541
δ	0.638	0.579	0.132	0.353	0.386	0.659	0.615	0.252	0.801	0.001	0.869
q	56	23	2	17	35	22	68	9	91	2	68

Table 4: Various metrics on latency arbitrage opportunities on a by-exchange basis for the 495 S&P 500 tickers, across the U.S. equities exchanges. Since I define latency arbitrage across a pair of exchanges, each arbitrage opportunity is double-counted as it is reported for each exchange involved the arbitrage. N is the mean daily number of arbitrage opportunities per ticker; T is the median of the daily mean duration per arbitrage opportunity in seconds; π is the mean profit per arbitrage opportunity in dollars; π_{total} is the mean daily profit per ticker; δ is the mean size (difference between the crossed BID and $OFFER$ on the two exchanges) per arbitrage opportunity in dollars; q is the mean quantity per arbitrage opportunity in number of shares.

	BZX	BYX	CHX	EDGA	EDGX	BX	NASD	PSX	ARCA	AMEX	NYSE
N	0.5	0.2	0.3	0.1	0.4	0.1	0.9	0.3	1.0	0.5	0.4
T	0.27	0.01	0.68	0.00002	0.06	0.01	0.5	0.02	0.34	0.01	0.12
π	0.25	0.13	0.08	0.07	0.24	0.12	0.47	0.06	0.68	0.06	0.36
π_{total}	3	2	4	1	3	2	6	0.3	11	9	13
δ	0.008	0.003	0.001	0.002	0.005	0.004	0.014	0.003	0.011	0.003	0.004
q	8.3	2.5	1.9	2.2	7.6	1.9	13.7	1.4	14.7	1.1	4.7

Table 5: Various metrics on latency arbitrage opportunities on a by-exchange basis for the 46 Russell 2000 tickers, across all exchanges. Data presented is as for Table 4.

5. Conclusion

This paper examined latency arbitrage opportunities that arise due to the fragmentation of trading across multiple stock exchanges. I analyzed order and quote data from consolidated and exchange market data feeds, which I accessed through MIDAS, a system available at the U.S. Securities and Exchange Commission. This work represents the first empirical study of the prevalence and profitability of latency arbitrage opportunities in over 500 ticker symbols across a representative set of U.S. stock exchanges. My results demonstrate that high-frequency traders employing a latency arbitrage strategy have numerous opportunities to realize trading gains. Based on my results, total potential profit in 2014 from these arbitrage opportunities is approximately $3.03 billion. The number and total potential profit of latency arbitrage opportunities varies by exchange. More opportunities are present on ARCA and NYSE, and in general there are more opportunities for cross-market latency arbitrage in larger market-capitalization stocks.

This work sheds light onto the extent of latency arbitrage opportunities in the market, but actual realized profits accumulated by high-frequency latency arbitrageurs cannot be determined solely from MIDAS data. Since MIDAS does not include identifying information, it is not possible to trace executed trades on two exchanges to a single firm. Therefore the dataset in this study cannot be used to ascertain whether or not HFTs are seeking to capitalize on these arbitrage opportunities, and if so, how successful they are in employing their latency advantages to do so. In addition, the dataset does not include off-exchange activity, so any latency arbitrage opportunities between exchanges are dark pools cannot be accounted for in this analysis.

References

Adler, J. (2012). Raging bulls: How Wall Street got addicted to light-speed trading. *Wired Magazine*.

Angel, J. J. and McCabe, D. (2013). Fairness in financial markets: The case of high frequency trading. *Journal of Business Ethics*, 112(4):5851–595.

Arnuk, S. and Saluzzi, J. (2009). Latency arbitrage: The real power behind predatory high frequency trading. Technical report.

Arnuk, S. L. and Saluzzi, J. C. (2012). *Broken Markets: How High Frequency Trading and Predatory Practices on Wall Street are Destroying Investor Confidence and Your Portfolio*. FT Press.

Baron, M., Brogaard, J., and Kirilenko, A. (2012). The trading profits of high frequency traders. Technical report, Commodity Futures Trading Commission.

BATS Global Markets (2015a). BATS Exchange: Alerts. `http://www.batstrading.com/alerts/`.

BATS Global Markets (2015b). Market volume summary. `https://www.batstrading.com/market_summary/`.

Biais, B. and Foucault, T. (2014). HFT and market quality. *Bankers, Markets & Investors*, 128:5–19.

Blume, M. E. (2007). Competition and fragmentation in the equity markets: The effect of Regulation NMS. *SSRN Electronic Journal*, pages 1–18.

Brogaard, J. (2010). High frequency trading and its impact on market quality. *Northwestern University Kellogg School of Management Working Paper*.

Budish, E., Cramton, P., and Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4):1547–1621.

CBOE Stock Exchange (2014). Regulatory Circular RG14-046.

Chicago Stock Exchange (2015). System status archive. `http://www.chx.com/system-status/archive/`.

Cohen, S. N. and Szpruch, L. (2012). A limit order book model for latency arbitrage. *Mathematics and Financial Economics*, 6:211–227.

Ding, S., Hannah, J., and Hendershott, T. (2014). How slow is the NBBO? A comparison with direct exchange feeds. *Financial Review*, 49(2):313–332.

Gaffen, D. and Curran, R. (2012). Measuring arbitrage in milliseconds. *The Wall Street Journal*.

Goldstein, M. A., Kumar, P., and Graves, F. C. (2014). Computerized and high-frequency trading. *The Financial Review*, 49(2):177–202.

Hasbrouck, J. (2010). The best Bid and Offer: A short note on programs and practices. *SSRN Electronic Journal*, pages 1–19.

IEX Group (2015). Letter to Brent J. Fields, Secretary, SEC, from Sophia Lee, General Counsel, IEX (November 23, 2015).

Jarrow, R. A. and Protter, P. (2012). A dysfunctional role of high frequency trading in electronic markets. *International Journal of Theoretical and Applied Finance*, 15(3).

Levine, M. (2014). Goldman Sachs got lost in its own dark pool. *Bloomberg*.

Lewis, M. (2014). *Flash Boys: A Wall Street revolt*. W. W. Norton & Company.

Narang, R. K. (2014). Exposing the falsehood of a prominent HFT critic's arguments. http://www.cnbc.com/id/101668593.

Nasdaq (2015). Nasdaq market system status search. http://www.nasdaqtrader.com/Trader.aspx?id=MarketSystemStatusSearch.

National Stock Exchange (2015). The National Stock Exchange re-launches trading operations. http://www.nsx.com/index.php/news-views/414-the-national-stock-exchange-re-launches-trading-operations.

NYSE (2015). Market status history. https://www.nyse.com/market-status/history.

O'Hara, M. and Ye, M. (2011). Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3):459–474.

Patterson, S. (2014). High-speed stock traders turn to laser beams. *The Wall Street Journal*.

Popper, N. and Protess, B. (2013). To regulate rapid traders, S.E.C. turns to one of them. *The New York Times*.

Schneider, D. (2012). The microsecond market. *IEEE Spectrum*, pages 66–81.

Stone, G. (2014). SIP and feeds. Technical report.

U.S. Department of the Treasury, Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, U.S. Securities and Exchange Commission, and U.S. Commodity Futures Trading Commission (2015). Joint staff report: The U.S. Treasury Market on October 15, 2014.

U.S. Securities and Exchange Commission. Self-regulatory organizations; edgx exchange, inc.; notice of filing and immediate effectiveness of a proposed rule change to clarify the use of certain data feeds. Release No. 3474072; File No. SREDGX 201502.

U.S. Securities and Exchange Commission (2005). Regulation NMS. 17 CFR Parts 200, 201, 230, 240, 242, 249, 270.

U.S. Securities and Exchange Commission (2013). Market information data analytics system. http://www.sec.gov/marketstructure/midas.html.

U.S. Securities and Exchange Commission (2014a). Order book reporting methods and their impact on some market activity measures. http://www.sec.gov/marketstructure/research/highlight-2014-03.html#.VkS-k7czbGg. Data Highlight 2014-03.

U.S. Securities and Exchange Commission (2014b). Self-regulatory organizations; national stock exchange, inc.; notice of filing and immediate effectiveness of proposed rule change to cease trading on its trading system. Release No. 34-72107; File No. SR-NSX-2014-14).

U.S. Securities and Exchange Commission (2015). Rule 611 of regulation nms. Memo to SEC Market Structure Advisory Committee.

Wah, E. and Wellman, M. P. (2013). Latency arbitrage, market fragmentation, and efficiency: A two-market model. In *14th ACM Conference on Electronic Commerce*, pages 855–872.

Wheatley, M. (2010). We need rules to limit the risks of superfast trades. *Financial Times*.