ESSAYS ON THE U.S. EQUITY SPEED BUMP AND NATIONAL MARKET SYSTEM

Jueheng Zhu

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy in Finance

Tepper School of Business

Carnegie Mellon University

Dissertation Committee:

Chester S. Spatt (Co-Chair)
Duane J. Seppi (Co-Chair)
Burton Hollifield
Ariel Zetlin-Jones

ACKNOWLEDGEMENT

I am deeply indebted to the unreserved support and guidance of my dissertation co-chairs, Chester Spatt and Duane Seppi. I am sincerely grateful to members of my dissertation committee, Burton Hollifield and Ariel Zetlin-Jones, for their insightful feedback.

I thank Chris Telmer and Bryan Routledge for their support and commitment to helping finance PhD students.

I thank all the finance faculty at Carnegie Mellon for their advice and inspiration. I want to express my gratitude to Lawrence Rapp and Laila Lee for their incredible administrative support.

ABSTRACT

In the U.S. equity market, high-frequency trading firms with access to fast networks engage in latency arbitrage. In this dissertation, we study both theoretically and empirically whether the artificial latency ("speed bump") implemented by Investors Exchange (IEX) and NYSE American protects slow traders from HFT predatory trading and improves market quality.

In the first chapter, we look at daily TAQ data around the implementation of these two speed bumps, and find empirical evidence that the IEX speed bump improves the overall market quality, while the NYSE American speed bump makes little difference. We find that the IEX speed bump has a larger impact on more volatile stocks and stocks traded on exchanges closer to IEX, while the NYSE American speed bump mostly affects trading handled by its Designated Market Makers. We also study latency reductions on several NYSE exchanges and find that they do not improve market quality.

In the second chapter, we build a model to study how a speed bump affects investors and the market. Fundamental investors, high-frequency arbitrageurs and market makers trade on two exchanges. When we introduce a speed bump to one of them, overall price discovery improves while the market has lower liquidity. Fundamental investors become more profitable at the expense of HF traders, while uninformed investors can be better or worse. We find that investors' trading behavior depends on information production, communication between market makers, as well as venue choice of uninformed investors, which in turn determine the impact of introducing a speed bump.

TABLE OF CONTENTS

Chapter 1

U.S. equity speed bumps and National Market System

1 Introduction

In asset trading, latency is defined as the lag between an input and an output. For example, the order execution latency is the time it takes for an exchange to execute an order after its submission. The transmission latency is the time it takes for a message to be sent from one place to another. The National Best Bid & Offer (NBBO) latency is the time it takes for the Securities Information Processing unit (SIP) to update the NBBO price once a new best price is displayed on any public exchange. Thanks to the development of technology, latency in the U.S. capital market has been greatly reduced. In terms of equity trading, latencies are now on the order of microseconds (one millionth of a second/one thousandth of a millisecond). Since trading is happening faster, a group of high-frequency trading (HFT) firms with superior speed technology invest in high-speed networks and develop computerized algorithms to capture arbitrage opportunities in the tiniest possible time-frame. In this paper we study one type of HFT activity: latency arbitrage, and one trading rule innovation called "speed bump" that aims to deter it.

Latency arbitrageurs, sometimes referred to as "front runners" or "quote snipers", use algorithms to pick off stale quotes in the market or to reprice their own quotes by being ahead in the queue of orders. The success of this type of arbitrage depends on being faster (in other words, having lower latency) when transmitting messages among exchanges, submitting and processing orders, gathering information on the

latest NBBO price, or a combination of all of the above. A speed bump is an artificial delay of message transmission from and to an exchange, intended to neutralize the speed advantages of latency arbitrageurs.

There are longstanding controversies about latency arbitrage. Opponents claim that latency arbitrage harms market quality, brings about a speed technology "arms race" among HFT firms that wastes social resources, and allows exchanges to have monopolistic power over data pricing.

Could a speed bump address the issues above? This paper uses daily TAQ data to test empirically the impact of inclusion of the IEX speed bump into National Market System. Using the IEX phase-in procedure as a natural experiment, this paper runs an unconditional regression that shows that for all stocks, the IEX speed bump improves the aggregate liquidity, encourages price discovery and discourages quote sniping both inside IEX and in the aggregate market, while not deterring the overall trading activity. In addition, this paper finds that the market quality improvements are more statistically significant for stocks that are heavily traded on the IEX. This paper then looks at the cross-sectional variations of stocks and transactions, and finds that more volatile stocks and transactions happening closer to the IEX or the central reporting SIP servers enjoy more market quality improvements after the implementation of the IEX speed bump

This paper also looks into the impact of the NYSE American speed bump as well, and finds that, although it is very similar to the IEX speed bump, it has very different impact due to a few key design distinctions.

The speed bump design was first introduced by the Investors' Exchange (IEX) in 2013. Specifically, all inbound and outbound trading messages are delayed by 350 microseconds. The concept of a "speed bump" became much more well-known since the SEC approved IEX operating as a public stock exchange with that design. In the SEC press release on June 17th, 2016,[1] the SEC said it approved IEX's application to register a national exchange because the Commission believes this action would "promote competition and innovation," so that the market could "continue to deliver robust, efficient service to both retail and institutional investors."

[1] https://www.sec.gov/news/pressrelease/2016-123.html

There's been much debate on whether a speed bump will help IEX and the market achieve these goals. Some research suggests that could be the case. Shkilko & Sokolov (2020) find empirical evidence that a brief halt of latency arbitrageurs' speed advantages is associated with lower adverse selection and lower trading costs. Gonçalves, Kräussl & Levin (2019) examine the implementation of a speed bump on a specific order type M-ELO on Nasdaq, and find that it decreases the average number of mini price crashes, while improving liquidity provision. In an experimental setup, Khapko & Zoican (2019) find that certain speed bump designs reduce investments in speed. However, Prof. Charles M. Jones's comment letter to the SEC[2] (March 2, 2016) expressed concern about IEX running as a public exchange. He argued that such a speed bump, if implemented by an exchange, would be far from de minimis in trading and could potentially harm market liquidity. The mutual fund AQR Capital Management argued that a speed bump to deter latency arbitrageurs could discourage the market makers as well.[3]

Why focus on a speed bump of the IEX, an exchange that has less than 3% of the trading volume in U.S. equity market? Regulation NMS requires that the NBBO price must consider quotes on all public exchanges, and thus the impact of a speed bump will not stay inside IEX. Any U.S. equity investor currently is in a market where the best prices (NBBO) are jointly determined by quotes from multiple linked exchanges with or without a speed bump.[4] Thus, it is important to understand the impact of a speed bump on the microstructure of equity trading.

A speed bump is another layer of intricacy in the already complicated U.S. equity trading microstructure. Thus, we not only study the aggregate impact of the IEX speed bump, but attempt to disentangle different aspects of the impact of a speed bump.

Stock characteristics. Theoretical research suggests that cross-sectional variations of stocks might affect the impact of a speed bump. Gonçalves, Kräussl & Levin (2019) find that the improvement on market stability (fewer mini-crashes) and liquidity provision (less algorithmic trading) from a specific order type with a speed bump varies depending on stock characteristics. Zhu (2019) builds a model of two exchanges

[2]https://www.sec.gov/comments/10-222/10222-433.pdf

[3]https://www.wsj.com/articles/high-frequency-hyperbole-1396394601

[4]Li, Ye & Zheng (2021) noticed the proliferation of orders types on stock exchanges to bypass Reg NMS and keep orders from being routing out to other exchanges. However, there's still large portion (43%) of orders that's routable.

where both informed traders and high-frequency arbitrageurs trade, predicting that a speed bump in general improves price discovery at the cost of liquidity, and the magnitude of improvement is larger for stocks with higher volatility. Aoyagi (2020) finds that although a speed bump helps liquidity provision, it lowers the marginal cost of being faster and encourages equilibrium speed acquisition. When the information variation of a stock is too small, this backfire effect on speed acquisition dominates the market quality improvement of a speed bump, leading to ambiguous empirical results about the impact of a speed bump.

This paper uses the diff-in-diff method to test cross-sectional variations after a speed bump being introduced in the market. We find that the market quality of large, liquid stocks with higher information variation improves more under the IEX speed bump.

Relative distance The matching engines of 13 U.S. public equity exchanges[5] are in different geographical locations. But what matters is the relative distance between each other, and the relative distance to the SIP. Technically speaking, the relative distance between the direct data feed server of the exchange and algorithm trading server of HFT latency arbitrageurs is an important factor as well. However, since almost every successful latency arbitrageur put their server next to the exchange proprietary server ("co-location"), that distance is infinitely small and can be ignored. Using daily TAQ data we find that the cross-market market quality impact of a speed bump is indeed a function of the relative distance of the stock trading: trading that happens on an exchange that is closer (has lower latency) to the SIP or IEX will be more affected by the speed bump of IEX.

Implementation A symmetric speed bump slows down everyone. An asymmetric speed bump only slows down a certain group. Thus far, the SEC has approved two symmetric speed bumps: the IEX speed bump on 2016 and the NSYE American speed bump in 2017. However, the SEC has been cautious in approving any form of an asymmetric speed bump. In February 2020, the SEC disapproved a plan from CBOE to introduce a four-millisecond speed bump to EDGA, one of its exchanges. The plan was intended to benefit market markers because cancel messages wouldn't be subject to the delay. The regulator called CBOE's proposal "discriminatory" and said that CBOE had not provided enough evidence to show that

[5]There were 13 exchanges within U.S. National Market System during our research period, 2016-2018. By 2022 there are 16 exchanges.

it would benefit the markets by curbing ultrafast trading strategies.[6] The Committee on Capital Markets Regulation objected as well, stating that asymmetric speed bumps confer an advantage on a select group of traders.[7] Similar concerns have been raised when the Canadian exchange TSX Alpha implemented a speed bump, but allowed investors to pay for a specific type of order that bypassed the artificial delay. Chen, Foley, Goldstein & Ruf (2017) find that the asymmetric speed bump segments order flow and increases profits for fast liquidity providers at the expense of other liquidity providers; the aggregate market quality decreases. In a laboratory market, Khapko & Zoican (2019) find that an asymmetric speed bump reduces investments in speed more than does a symmetric speed bump, and that a larger magnitude of the speed bump leads to lower investments in speed. These research suggest that even a tiny variation in the speed bump implementation could lead to drastically different results.

We compare the NYSE American speed bump with the IEX speed bump, two very similar designs with several subtle distinctions. IEX once expressed concern that this seemingly identical speed bump design might not work for NYSE American because the geographical locations of NYSE American and IEX are different. We find that, unlike IEX, the NYSE American speed bump does not cause any significant change in price discovery or liquidity, neither inside NYSE American nor on other exchanges. In addition, we find that Designated Market Makers (DMM) of NYSE American, i.e., specialists that commit to help facilitate the trading for certain stocks, benefit most from the NYSE American speed bump. This points out the regulatory risk that seemingly similar speed bump could still grant differential advantages to a certain group.

After investigating two U.S. equity speed bumps, we ask a broader question: Does lower latency produce better market quality? The transition from traditional broker-dealer trading to electronic matching caused a huge latency reduction (from seconds to milliseconds) for the U.S. equity market, and came with an improvement of overall market quality. For example, Hasbrouck (1995) discovers that switching to electronic trading encouraged more prompt price discovery as the general predictability of the equity market fell. In the recent decades, latency has been further reduced from milliseconds to microseconds thanks to the development of speed technology. Researchers have debated over whether it has positive impact on market

[6] https://www.sec.gov/rules/sro/cboeedga/2020/34-88261.pdf
[7] https://www.capmktsreg.org/wp-content/uploads/2019/12/Nothing-But-The-Facts-Asymmetric-Speed-Bumps.pdf

quality. For example, while Hendershott, Jones & Menkveld (2011) shows that low-latency algorithm trading improves liquidity by significantly reducing the bid-ask spread, Aquilina, Budish & O'Neill (2022) finds no evidence of liquidity improvements for any millisecond-level latency reduction.

This paper uses daily TAQ data to test empirically the impact of NYSE Arca's and NYSE National's migrations to NYSE's new trading platform, Pillar, where new technologies further reduce the latency. We find that this latency reduction does not change price discovery or liquidity significantly, while slightly encouraging HFT activity.

This paper is related to three branches of literature: The study of speed innovations in a fragmented market; research on latency arbitrage; and both theoretical and empirical investigation on the impact of a speed bump.

Several papers study speed innovations in a fragmented market setting. Angel, Harris & Spatt (2015) find that slow traders are more vulnerable in fragmented markets where quote-matchers can front-run orders in one market by trading in another market. Wang (2018) explores how and why exchanges compete on order processing speeds as a service appealing to fast traders who take advantage of the Order Protection Rule. Lee (2019) uses the Kyle model to explore how a symmetric cross-venue latency affects informed traders' optimal strategy, where a larger latency makes cross-venue order-flow less informative. This paper shows that after incorporating the IEX speed bump into National Market, liquidity and price discovery improves; however, when a seemingly identical speed bump is implemented on NYSE American (which has a different artificial latency structure), aggregate market quality does not change. This finding echoes the importance of considering the network when describing or modeling the microstructure of the U.S. equity market.

To estimate the profitability of latency arbitrage, Aquilina, Budish & O'Neill (2022) examine message data from 43 trading days in August to October 2015 at the London Stock Exchange. They find that latency arbitrage races are very frequent (one per minute for FTSE 100 stocks), extremely fast (the modal race lasts 5-10 millionths of a second), and account for a large portion of overall trading volume (about 20%). Race participation is concentrated, with the top-3 firms accounting for over half of all race wins and losses. Shkilko & Sokolov (2020) study a series of exogenous weather episodes that temporarily remove the speed advantages

of the fastest traders by disrupting their microwave networks, and find that those disruptions are associated with lower adverse selection and lower trading costs. They also show under a theoretical framework that a long term removal of speed differentials results in similar effects and also increases gains-from-trade. Our results confirm that a speed bump helps the market quality, and that it is a feasible technology innovation alternative to the traditional, more direct regulatory intervention.

There are a few theoretical studies on the impact of a speed bump. Aldrich & Friedman (2017) compare the execution performance of three order types in particular: midpoint peg, primary peg and discretionary peg,[8] the latter two of which are unique to IEX. Their model predicts that the IEX speed bump will generally improve price efficiency and lower transactions cost while increasing delay costs. Zhu (2019) builds a model of two exchanges where both informed traders and high-frequency arbitrageurs trade, predicting that a speed bump in general improves price discovery at the cost of liquidity. In addition, different distances between two exchanges or the variations of volatility of stocks would drive different results. Aoyagi (2020) finds that a speed bump could lower the marginal cost of being faster and encourages the equilibrium speed acquisition. By investigating when this backfire effect of speed acquisition dominates the market quality improvement of a speed bump, the paper offers one explanation for ambiguous empirical results on the impact of a speed bump. Our results show that there are indeed conflicting forces within the aggregate impact of a speed bump, and that the design needs to be very careful to achieve its intended goal.

Using data from Betfair, a horse-racing betting exchange, Brown & Yang (2016) find evidence that the speed bump there protects slower traders and that fast traders develop strategies to circumvent the speed bump. Gonçalves, Kräussl & Levin (2019) study high-frequency order book message data around the implementation date of Midpoint Extended Life Order (M-ELO) on Nasdaq. M-ELO is a marketable order that tracks NBBO mid-point, is subject to a 1/2 second speed bump, and can only execute against the same order type of opposite direction[9]. They find that the introduction of the M-ELO decreases the average number of mini-crashes while increases liquidity provision. Using TAQ data, Hu (2019) finds that

[8]Midpoint peg is a marketable order pegged to the NBBO mid-price. Primary peg is a marketable order pegged to NBBO that has discretion to execute at a price equal to or better than best bid/ask. Discretionary peg is a marketable order pegged to NBBO that has discretion to execute at a price equal to or better than NBBO mid-point. See more at https://iextrading.com/trading/order-types/

[9]https://www.nasdaq.com/solutions/midpoint-extended-life-order-m-elo

the introduction of a speed bump in IEX helps the overall price discovery and liquidity in the market. The aggregate impact in our test agrees with the results from Hu (2019), but with a more detailed decomposition of the impact: price discovery and liquidity could potentially change in opposite directions inside and outside the IEX; the impact on individual stocks is heterogeneous depending on characteristics and geographical locations of trading servers. In addition, by comparing the results from IEX to NYSE American, this paper shows that not all designs of speed bump achieve the same impact.

The rest of this paper proceeds as follows. Section II introduces the institutional background of latency arbitrage and speed bumps in the U.S. equity market. Section III describes the data and market quality measures we use. Section IV shows the results for the IEX speed bump. Section V shows the results for NYSE American speed bump and compares it to the IEX speed bump. Section VI shows the results for NYSE Arca and NYSE National migration to Pillar. Section VII concludes.

2 Institutional background

In this section, relevant institutional background is provided, such as the trading protocol of Reg NMS, the definition of the NBBO, the controversies of latency arbitrage, implementation of two U.S. equity exchange speed bumps, and a trading technology upgrade that leads to latency reduction within NYSE exchanges.

2.1 Reg NMS

The public U.S. equity market consists of a group of national exchanges that are interconnected and follow similar rules, the Regulation National Market System (Reg NMS).[10] Table 1 shows the market shares of all public exchanges [11] using three measures: dollar amount, number of trades, and total volume. Rankings of market shares vary slightly under different measures, but Nasdaq and NYSE are dominant. Transactions not on exchanges (dark pools, Alternative Trading Systems, etc.) reported by FINRA Trade Reporting

[10] https://www.sec.gov/fast-answers/divisionsmarketregmrexchangesshtml.html
[11] https://www.nasdaqtrader.com/trader.aspx?id=FullVolumeSummary#

Facility (TRF) are not the subject of this paper, but have their unique information content: Ernst, Sokobin & Spatt (2022) find that the publication of these off-exchange transactions leads to a sharp burst in trading and quoting activity, suggesting that market participants learn from those reports.

Table 1: U.S. Exchange equity market shares

Exchange	$ Amount %	Trades %	Volume %
NASDAQ	19.1	23.6	16.6
NYSE	11.7	9.8	12.7
ARCA	9.3	8.9	8.2
BATS Z	6.8	8.8	5.8
EDGX	5.3	7.1	5.6
IEX	3	5.3	3.3
BATS Y	2.5	4.3	2.8
EDGA	2.1	3.8	2.2
BX	1.3	2	1.3
PSX	0.7	0.8	0.6
AMEX	0.2	0.3	0.3
CBOE	0	0	0
MWSE	0	0	0
NSX	0	0	0
Public Exchange total	62.0	74.7	59.4
FINRA TRF	38.0	26.3	40.6
Total	100	100	100

Data is retrieved through Nasdaq in Oct 2019. $ Amount is total dollar value traded. Trades is total number of trades. Volume is the total number of shares traded.

Reg NMS requires that all exchanges both report to and honor trades and quotes from the Securities Information Processor (the SIP), a public central server that receives and disseminates information. The best quotes available in the market is called National Best Bid & Offer (NBBO). Rule 611 protects the best automated quotes of exchanges by obligating other venues to not execute trades at inferior prices. If a venue has inferior quotes, then it may cancel, post, or route incoming orders to exchanges with better prices. Since Rule 611 prevents "trading-through" the best quotes it is often called the "order protection rule," or "no trade-through" rule.

2.2 HFT Latency arbitrage controversies

Whenever a price shock happens anywhere in the market, the HFT traders can observe it and rush to other exchanges before the NBBO updates. The HFT traders are then able to snipe the stale quote, almost immediately unload the position and make a profit.

Opponents claim that latency arbitrage harms market quality. First, they argue that quote sniping may worsen the adverse selection problem faced by market makers, forcing them to widen the spread. In a U.K. Financial Conduct Authority report, Aquilina, Budish & O'Neill (2022) examine message data from London Stock Exchange and estimate HFT traders earn nearly $5 billion on global stock markets in 2018 by taking advantage of slightly out-of-date prices, imposing a small but significant tax (0.0042% of daily stock-trading volume) on investors. Second, the limit order book could be strategically manipulated by HFT to scalp profits from slow traders (Cumming, Johan & Li (2011)). Third, investors may experience "phantom liquidity" where displayed quotes they attempted to access are moved away from the inside the moment they place an order. Chung & Chuwonganant (2014) document that during periods of uncertainty, HFT market makers suddenly withdraw liquidity, and sometimes even demand liquidity instead. In the most extreme case, algorithms trigger each other and a series of cancellation might evolve into a flash-crash, where almost all liquidity is pulled out of the market in an instant, causing panic and disorder.

Latency arbitrage is also blamed for bringing about a transmission speed technology "arms race" among HFT firms. Since every U.S. national equity exchange keeps a continuous double auctions (CDA) limit order book, even an extremely small fraction of a second provides the faster trader time priority. Budish, Cramton & Shim (2015) argues that the huge amount of resources HFT firms spend on speed technology arms race is a cost borne by other market participants. Michael Lewis's *Flash Boys* documents Spread Network's $300 million investment in fiber optics from the futures market in Chicago to stock exchanges in New York to cut transmission time by 3 milliseconds (3 thousandths of a second). Thanks to the development of microwave technology which reduces the transmission time further, Spread Network's optics became obsolete by its completion, and the company was sold for $131 million in 2019.[12]

[12]Prof. Paul Krugman commented that huge sum was given to this project while receiving little or nothing. see https://www.nytimes.com/2014/04/14/opinion/krugman-three-expensive-milliseconds.html

A somewhat unexpected consequence of latency arbitrage is monopolistic data pricing by exchanges. Efficient sniping algorithms and superior speed of transmission are not enough for a latency arbitrageur; co-location with the exchange servers and subscription to a direct data feed from the exchanges are essential. A direct feed contains a real-time data stream of trade and quote information. These data are then compiled by exchanges and published at the end of day as "daily TAQ", the dataset this paper uses. But more importantly, a direct feed contains in-depth order book dynamics known only by an exchange and given exclusively to the direct feed subscribers. In 2019, the New York Stock Exchange (NYSE) proposed a plan to construct a pair of antennas designed to shave two millionths of a second off the time it takes for high-frequency traders to access its computer systems.[13] Several HFT firms expressed strong objection. They were afraid this tiny latency improvement would force them to accept very high NYSE pricing for the subscription, because otherwise they would be too disadvantaged compared to subscribers. In a recent study on the cost of exchange services,[14] researchers from IEX made a similar claim that the importance of speed enables exchanges to extract monopolistic rents for latency reducing products. CME Group built a wireless tower just outside its trading center in 2018 and intended to sell the access to the 35 dishes on that tower that can transmit signals faster than anyone from CME to data centers in New York. The CME tower would have served similar functionality as NYSE's antenna, but it has not been used yet. A company called Scientel has been trying to build an almost identical wireless tower just a few yards away. CME and Scientel has been fighting aggressively over the legitimacy of Scientel tower.[15] This demonstrates again the importance of relative speed advantage (however small) in the world of HFT trading.

2.3 Alternative measures against latency arbitrage

Incorporating a speed bump into an exchange is not the only attempt to deter latency arbitrage in the U.S. equity market. Politicians in both Europe[16] and the U.S.[17] have pushed for a financial-transaction tax, a policy aimed in part at curbing high-speed trading. However, the specific tax rate is very hard to

[13] https://www.wsj.com/articles/nyse-antennas-spark-high-speed-trader-backlash-11565272102
[14] https://iextrading.com/docs/The%20Cost%20of%20Exchange%20Services.pdf
[15] https://www.bloomberg.com/news/features/2019-03-08/the-gazillion-dollar-standoff-over-two-high-frequency-trading-towers
[16] https://www.wsj.com/articles/germany-pushes-forward-on-european-financial-transactions-tax-11576074482
[17] https://www.wsj.com/articles/democrats-aim-for-financial-transactions-tax-11551818240

determine; a large part of latency trading is rapid order canceling, which would not be subject to tax. Thus far there has been no implementation of a HFT-targeted transaction tax or SEC high-speed trading fee in the U.S. equity market.

Budish, Cramton & Shim (2015) have proposed an alternative trading protocol to address this issue, a "frequent batch auction (FBA)," where orders are processed discretely (e.g. every tenth of a second). Aldrich & Vargas (2019) implement a laboratory financial market to contrast the performance of FBA against continuous double auction (CDA). Their evidence suggests that, relative to the CDA, the FBA exhibits less predatory trading behavior, lower investments in speed technology, lower transaction costs, and lower volatility in market spreads and liquidity. In 2018, IntelligentCross was launched as an Alternative Trading System (ATS) that matches mid-point orders every millisecond and limit orders every few hundred microseconds.[18] A discrete auction is very distinct from the prevalent CDA protocol, thus is not yet incorporated into the National Market System (NMS).

2.4 Speed bumps in the U.S. equity market

Speed bumps have been introduced in equity, foreign exchange, as well as futures markets, in both North America and in Europe, as is shown in the following time-line:[19]

[18] The specific intervals between each batch of trading are determined by the platform's proprietary algorithms. See more at https://imperativex.com/intelligentcross/

[19] https://www.wsj.com/articles/more-exchanges-add-speed-bumps-defying-high-frequency-traders-11564401611

Figure 1: Different adaptations of Speed bumps

Among them, IEX, NYSE American, Nasdaq, IntelligentCross and CBOE EDGA trade U.S. equity. Nasdaq did not apply a speed bump to all stocks, but only to a specific order type called Midpoint Extended Life Orders. IntelligentCross is an Alternative Trading System (ATS) that employs discrete matching rather than artificial delay. CBOE EDGA's proposal of an asymmetric speed bump was rejected by the SEC on February 2020.[20]

IEX was the first U.S. equity exchange to adopt a speed bump. IEX began as a dark pool Alternative Trading System (ATS) in Oct 2013. In April 2015, IEX introduced displayed trades and top-of-book quotes (the TOPS system) and started to operate as a lit Electronic Communications Network (ECN). The SEC approved IEX's application to become a national exchange in June 2016. IEX is known for its 350-microsecond

[20] https://www.sec.gov/rules/sro/cboeedga/2020/34-88261.pdf

speed bump, implemented with a shoe box containing 38 miles of coiled fiber. Because IEX's matching engine is only accessible through the coil, all incoming and outgoing messages have to travel an additional 38 miles, which creates its 350 microsecond access delay.

On July 24, 2017, NYSE American debuted its own speed bump. The speed bump is very similar to the one implemented by IEX, but with four distinct differences. First, although many parameters are the same, the implementation can differ from the IEX speed bump since the IEX and NYSE American have data servers located in different places. Second, the fee structure of NYSE American is maker-taker, while the IEX does not pay any rebates to market making. Third, NYSE American allows for co-location, which effectively allows the HFT to bypass the speed bump if they choose to purchase the service. Last, NYSE American has the system of Designated Market Makers, who are obliged to provide liquidity. Later in this paper, we will see that the empirical impact of a speed bump is different on IEX versus NYSE American. These four distinctions are potential reasons for that observation.

2.5 NYSE Pillar migration

Starting 2017, NYSE has been gradually moving its exchanges to a new platform called "Pillar." Pillar has new integrated trading technology that will enable member firms to connect to all of NYSE equities and options markets with improved efficiency and reduced latency. For example, some order types are standardized and can now travel from one NYSE exchange to another without modification; the information transmission among all NYSE exchanges are faster. To date, NYSE, NYSE American Equities, NYSE Arca Equities, NYSE National and NYSE Chicago have been migrated to NYSE Pillar.[21] In this paper, we study the migration of NYSE Arca (an exchange specialized in trading Exchange Traded Products (ETP)) and NYSE National (an exchange with no listing services) as two examples of this latency reduction effort by NYSE.

[21] https://www.nyse.com/pillar

3 Hypothesis development

This paper looks into four aspects of the impact of a speed bump: liquidity, price discovery, benefit to slow traders, and trading activity on the exchange with a speed bump. When a speed bump is implemented in one exchange, we call the impact on that exchange "local" and the impact on all exchanges "global". To develop the hypotheses tested in this paper, we make several assumptions:

A1 Competitive market-making connects exchanges in National Market System. We assume that pricing is competitive among exchanges. Market makers offer their best quotes that set or match the NBBO, such that orders submitted elsewhere are routed to them. The inclusion of the IEX in the National Market System means that the best quotes in IEX (and the speed bump that shapes the price discovery mechanism of those quotes) affect all other exchanges in the U.S. equity market.

A2 Aggregating all exchanges without a speed bump is equivalent to the market. This paper studies IEX and NYSE American, the only two national exchanges that have exchange-wide speed bumps. The total market share of these two exchanges was less than 5% as of our sample period. Thus, the impact of a speed bump on other exchanges will be a good proxy for the impact on the aggregate market (or vice versa). For example, we calculate the impact of IEX becoming a National Exchange on all other public exchanges, and compare the results with the impact on all exchanges (including IEX). Almost all results are not statistically different from one another, and thus we only show results of the aggregate market including IEX in later sections.

A3 Fundamental investors do not reverse position or cancel quotes quickly. In terms of ways of trading, we define a fundamental investor as someone who buys and holds/ sells and walks away (i.e. an "end user" who holds a non-zero position after a trading), while a sniper as someone who finds another counter-party to trade against immediately after the prior trading is done (in order to keep inventory risk as low as possible). In terms of ways of quoting, we define a fundamental investor as someone who posts a quote that stays on the order book for at least a second, while a sniper as someone who frequently submits/cancels

quote.[22] Since the data we have is anonymous, it is impossible to tell whether a trade goes to or whether a quote comes from a fundamental investor vs. a sniper. For the purpose of estimation, we assume that orders quickly reversed within a very short period of time go to arbitrageurs, and that all the other orders go to fundamental investors. We also assume that the cancel-to-trade ratio largely represents arbitrage activity from fast traders. Based on these two assumptions, we are able to detect whether a speed bump has heterogeneous impact on different types of investors by looking at how these two ratios (the cancel-to-trade ratio and the fundamental-to-sniper ratio) change before and after a speed bump.

A4 More order sniping results in negative price autocorrelation after large volume trading

Quote sniping algorithms are more likely to be triggered by large volumes of trading that send out stronger signals of the dynamics of supply and demand. Moreover, quote snipers, or arbitrageurs, do not keep their inventory for long and need to unload their positions fast. This reversal of trading directions would lead to a negatively correlated price series. Thus, we assume that a large volume predicts negative price autocorrelation better when there is more sniping.

3.1 Liquidity

We measure liquidity from two perspectives: displayed spreads (ex-ante liquidity) and price impact (ex-post liquidity).

H1.1 A speed bump improves local liquidity Market makers are faced with an adverse selection problem, where their stale quotes may be sniped by arbitrageurs. When market makers expect more snipers, they require a higher bid-ask spread to compensate for potential loss of being the victim of quote sniping. Similarly, when a stream of orders is sent to market makers, they may adjust their quotes more aggressively away from the execution price of earlier orders, fearing that they would otherwise leave too much money on the table by offering a not well-informed quote that is too good for snipers. Thus, if a speed bump

[22] It might not be obvious why quote snipers need to cancel orders and fall into this category, since the literature usually models sniping as fast market orders that takes up liquidity. The reason is that snipers build a net of traps consisting of small limit orders on all exchanges for all possible prices. Whenever a series of orders are executed, the sniping algorithm is triggered and sniping starts. The dynamically evolving limit order trap requires frequent intra-day canceling and resubmitting.

discourages fast trading, we should expect the spread and price impact to be smaller (assuming market making is competitive).

H1.2 A speed bump improves global liquidity When a speed bump is implemented on an exchange, other exchanges are not subject to the speed bump directly. However, Reg NMS requires orders to be routed to the exchange with a speed bump if it has better price. Since market makers there are faced with less severe adverse selection problem, they tend to offer better prices. This competition could in turn force market makers in other exchanges to price more competitively with lower spreads and less aggressive price adjustments.

3.2 Price discovery

We measure price discovery from two perspectives: matching & setting best price (ex-ante price discovery) and price efficiency (ex-post price discovery).

H2.1 A speed bump improves local price discovery A speed bump does not directly affect market makers' willingness to match or improve NBBO price. However, since their quotes are better protected, on average they could offer better quotes, resulting in a higher probability of their quotes being the best price. If informed investors are protected, they are willing to disclose more information while trading, which will be reflected in price.

H2.2 A speed bump improves global price discovery If a speed bump protects and improves quotes inside an exchange, market makers on other exchanges may match or improve quotes more aggressively to attract order flow. The formation of an efficient price might be slower due to the inbound and outbound information delay of an exchange with a speed bump. When the protection from the exchange with a speed bump is appealing enough, informed investors might concentrate there and disclose more information. If information can be efficiently conveyed from the exchange to the market, we expect the total amount of information in price to be larger.

3.3 Fast vs. slow investors

One of the purposes of a speed bump is to favor fundamental, slow trading over arbitrage HFT activity. We use order sniping & canceling to measure fast traders activity. We define fundamental ratio as the percentage of trading volume that does not quickly reverse divided by all trading volume, and use it to measure slow investors' activity.

H3.1 A speed bump discourages local HFT arbitrage activity and benefits local slow investors If a speed bump discourages fast trading, there should be less arbitrage activity in general. Since massive order canceling is one key feature of arbitrage trading, we expect the cancel-to-trade ratio to drop. In addition, less arbitrage means less position reversal after large volumes of trading trigger quote sniping algorithms. Thus, after large volumes of trading we should observe less negative price autocorrelation. Meanwhile, if a speed bump benefits slow investors, we should observe an increasing fundamental ratio.

H3.2 A speed bump discourages global HFT arbitrage activity and benefits global slow investors In general, we do not expect a speed bump on one exchange to directly impact the arbitrage activity on other exchanges. However, when a quote from that exchange is the quote that defines the new NBBO, the protection inside one exchange with a speed bump is extended to the market. This means the NBBO arbitrage becomes harder for quote snipers who must go through the artificial delay, while the fundamental investors outside get the benefit. Thus, when an exchange with a speed bump discovers the best price, we expect to see less arbitrage activity and larger fundamental ratio, even outside the exchange.

3.4 Trading activity

We test whether adding a speed bump makes an exchange more appealing to the investors by measuring the trading volume market share of that exchange.

H4 A speed bump increases the trading volume market share of that exchange A speed bump could potentially improve execution quality, which is desirable by institutional investors. Even if that

is not the case, adding an artificial delay might induce some structural change that is appealing to a specific group of investors (or more generally, a certain way of trading). For example, an institutional investor who wish to trade at market fair price is better off if a speed bump slows everything else down but allows NBBO mid-price to be updated promptly. A quote sniping hedge fund could benefit from a reasonably-priced order type that allows its users to bypass a speed bump and see the price change before anyone else does. A market maker trying to avoid adverse selection should welcome a speed bump that slows down liquidity takers but not liquidity providers. By attracting more order flow to the exchange we expect its market share to increase.

4 Data and trading activity measures

The primary source of data is the daily Trade and Quote (TAQ) database from NYSE.[23] Daily TAQ includes data of all national exchanges under the National Market System. It records local best (top of the book) quotes from all exchanges and global best (national best bid/offer, NBBO) quotes of the entire market. It also provides records of all trades. The timestamp of each daily TAQ is accurate to the microsecond. A total of 5 months of historical data is used: Aug 2016, Sept 2016, July 2017, Aug 2017 and May 2018, which covers roughly 10 trading days both before and after the following important events[24]:

1. Aug 19-Sep 2, 2016, when IEX became a national exchange
2. July 24, 2017, when NYSE American debuted its own speed bump
3. Aug 21, 2017, when NYSE Arca migrated to Pillar
4. May 21, 2018, when NYSE National migrated to Pillar

For ease of data processing, some of the historical IEX trades and top of the book quotes are downloaded

[23] https://www.nyse.com/publicdocs/nyse/data/Daily_TAQ_Client_Spec_v3.0d.pdf

[24] In the case of Aug 21, 2017, we use 10 trading days before and 9 trading days after the event. In the case of May 21, 2018, we use 10 trading days before and 8 trading days after the event, because May 28, 2018 was Memorial Day, a NYSE holiday. In a robustness test, we run the same regression based on the data from 8 days before and 8 days after the events. The results we get are very similar to what we show in the paper.

from publicly available TOPS feed on the IEX website.[25] Securities characteristics such as market cap, volatility & turnover are from CRSP.

4.1 Market activity & quality measures

Using the aforementioned data, we construct a panel that records trading performance for every publicly traded stocks, on each trading day within a time window when latency changes happened. Notice some of the measures are based on intra-day data. In those cases, a daily average for each stock is obtained first and then put into the panel of (stock, day) pairs. This paper then tests how market activity and market quality change after a speed bump based on that panel. The standard deviations in those regressions are clustered by day. This section discusses our measures of market activity and market quality, and how we use those measures to map the hypothesis into empirical tests.

4.1.1 Liquidity

To measure liquidity, we look at A) the displayed liquidity a typical investor sees in the market, and B) the actual cost if he trades. For displayed liquidity, we use two alternative definitions of spread, both of which are considered to be indicators of better liquidity if the spread is smaller.

The quoted spread is defined as the daily average bid-ask spread for a certain stock on a particular exchange, normalized by dividing the NBBO mid-price and measured in percentage points. For example, IEX quoted spread of stock AAPL on day t is:

$$\text{Quoted Spread} = \frac{\text{IEX Best Offer of AAPL on day } t - \text{IEX Best Bid of AAPL on day } t}{\text{NBBO mid-price if AAPL on day } t} \tag{1}$$

Bid-ask spread could be a constant and lack variations for very liquid stocks since their quoted spread would always be a penny, the minimal unit under the current system. One alternative is to follow Roll's measure (1984). Roll spread assumes the market is efficient, and is a straightforward way to calculate implicit

[25] https://iextrading.com/trading/market-data/

spread using only trade data:[26]

$$\text{Roll Spread} = \frac{2\sqrt{-cov(\Delta p_s, \Delta p_{s-\Delta s}) \text{ of AAPL}}}{\text{NBBO mid-price of AAPL at time } s} \tag{2}$$

To further decompose investors' cost to trade, we look at different aspects of it and utilize the following formula:

$$\text{Effective spread} = \text{price impact} + \text{realized spread} \tag{3}$$

or equivalently:

$$\text{Transaction cost} = \text{informational shock} + \text{market making profit} \tag{4}$$

The effective spread is the total cost of trade for an investor. A market with higher spread is less liquid for investors. Effective spread is defined as the daily average of the differences between the execution price at the time of trade and the fair price (NBBO mid-price) shortly before, for a certain stock on a particular exchange, normalized by dividing the NBBO mid-price and measured in percentage points. For example, IEX effective spread of stock AAPL on day t is:

$$\text{Effective spread} = \frac{\text{Execution price of AAPL at time } s - \text{NBBO mid-price of AAPL at time } (s - \Delta s)}{\text{NBBO mid-price of AAPL at time } s} \tag{5}$$

Here s stands for the intra-day time-stamp of a trade or quote reported to SIP, and Δs is the reporting interval of the SIP that publishes NBBO price to the market (1 millisecond). The NBBO price at time s incorporates information from any trade that has a reporting timestamp $\leq s$ (or equivalently, any trade that happens between time $s - \Delta s$ and time s).

Price impact is the value change caused by the informational shock. It is defined as the daily average

[26] In practice, the first-order serial covariance of price changes can be positive. It happens rarely ($\leq 0.04\%$) and the magnitude when the correlation is positive is much smaller than the magnitude when the correlation is negative ($\sim 5.6\%$). Thus, when the covariance is positive, for computing efficiency purposes, we use absolute value of the covariance instead:

$$\text{Roll Spread} = \frac{2\sqrt{|cov(\Delta p_s, \Delta p_{s-\Delta s})| \text{ of AAPL}}}{\text{NBBO mid-price of AAPL at time } s}$$

of the differences between the fair price (NBBO mid-price) shortly after the trade and the fair price shortly before, for a certain stock on a particular exchange, normalized by dividing NBBO mid-price and measured in percentage points. The market incorporates the information of a trade during the time lag between when the trade happens and when we measure the NBBO mid-price. The literature uses a range from a few seconds to a few minutes as time lags, where more recent papers tend to use shorter time lags (Chen, Foley, Goldstein & Ruf (2017), Gonçalves, Kräussl & Levin (2019), Shkilko & Sokolov (2020)). Here we use 5 seconds[27]. For example, the IEX price impact of stock AAPL on day t is

$$\text{Price impact} = \frac{\text{NBBO mid-price of AAPL at time } (s+5) - \text{NBBO mid-price of AAPL at time } (s-\Delta s)}{\text{NBBO mid-price of AAPL at time } s} \tag{6}$$

The realized spread is the profit earned by market makers. It is defined as the daily average of the differences between the execution price and the fair price (NBBO mid-price) shortly after the trade, for a certain stock on a particular exchange, normalized by dividing NBBO mid-price and measured in percentage points. For example, the IEX realized spread of stock AAPL on day t is

$$\text{Realized spread} = \frac{\text{Execution price of AAPL at time } s - \text{NBBO mid-price of AAPL at time } (s+5)}{\text{NBBO mid-price of AAPL at time } s} \tag{7}$$

4.1.2 Price discovery

We measure price discovery both before and after trading. To measure how good a certain exchange finds best price for its investors, we look into whether the local quotes are informationally as good as or even better than best quotes in the market.

Specifically, "*NBBO relevant*" measures how often can the local best quotes from a particular exchange keep up with the best price in the market (NBBO)[28] . It is defined as the daily percentage of seconds where the quote of a stock from an exchange matches the NBBO. For example, the IEX NBBO relevant of stock

[27]For robustness this paper also tries using 1 second time lag, and the results are similar

[28]There are cases where the local best quote on a particular exchange is even better than the NBBO price. In the daily TAQ data, these quotes will be labeled as "NBBO-defining," while quotes at NBBO are labeled as "NBBO-eligible." Here I do not distinguish between them, and will call both NBBO relevant.

AAPL on a given trading day t is:

$$\text{NBBO relevant} = \frac{\text{seconds when best price of AAPL locally at IEX matches NBBO at day } t}{\text{total seconds in a trading day at day } t} \tag{8}$$

Similarly, "*NBBO setting*" measures how often does a particular exchange's quotes improve the existing best price in the market, which in turn sets the new NBBO. It is defined as the daily percentage of updates where the improvements of NBBO come from a particular exchange. For example, the IEX NBBO setting of stock AAPL on a given trading day t is:

$$\text{NBBO setting} = \frac{\text{updates when best price of AAPL locally at IEX improves NBBO at day } t}{\text{total updates in a trading day at day } t} \tag{9}$$

We use autocorrelation of absolute returns as a measure of after-trading price discovery. If price discovery is better, the market should be more efficient in incorporating information into the price. Stock prices would be harder to predict and have lower correlation with its past performance, thus smaller autocorrelation for absolute returns. We use absolute return series correlation of every second and every minute within a trading day as short-lived and long-lived measures. Here we make the simplifying assumption that the definitions of "short-lived" and "long-lived" are the same for different stocks.

Short-lived autocorrelation is defined as the daily autocorrelation of NBBO mid-price absolute return every 1 second. For example, the short-lived autocorrelation of stock AAPL on a given trading day t is:

$$\text{Autocorr(1sec)} = corr(\text{NBBO mid-price absolute return of AAPL over time } (s, s+1), \forall s \in \text{day } t) \tag{10}$$

Similarly, long-lived autocorrelation of stock AAPL on a given trading day t is the daily autocorrelation of NBBO mid-price absolute returns every 60 seconds:

$$\text{Autocorr(60sec)} = corr(\text{NBBO mid-price absolute return of AAPL over time } (s, s+60), \forall s \in \text{day } t) \tag{11}$$

We follow Hasbrouck (1991,1993) to decompose the volatility of stock prices into two parts: the permanent impact σ_p that represents the information content that's revealed in trading; and the transient impact σ_t that represents liquidity shock, inventory control effects, and noise trading. We define information ratio (IR) as the ratio of permanent impact over transient impact, which is a value between 0 and 1. A larger ratio would indicate a more efficient market.

$$\text{Information ratio} = \frac{\sigma_p \text{ of NBBO mid-price of AAPL}}{\sigma_t \text{ of NBBO mid-price of AAPL}} \tag{12}$$

4.1.3 Fast vs. slow trading

Although we do not directly observe the distinction between fast and slow trading, we construct three proxies to measure fast trading activity: volume-autocorrelation predictability, the cancel-to-trade ratio and the fundamental ratio.

Quote sniping is more likely to happen after large volumes of trading when stronger signals of the dynamics of supply and demand are revealed. When fast traders observe such signals, they trade against stale quotes and quickly reverse their positions to minimize inventory risk. This reversal generates negatively correlated prices. Thus, as a proxy for fast traders' sniping activity, we look at how well a large trading volume at any time s predicts the per transaction price autocorelation in the next 1 second (from time s to $s+1$) for a particular exchange, or the millisecond NBBO mid-price autocorrelation in the next 1 second (from time s to $s+1$) for the aggregate market[29]. To measure this, we run the following regression:

$$\text{price autocorrelation within time } (s, s+1) \sim \beta \cdot \text{trading volume at time } s + \gamma \cdot \text{controls} + \epsilon \tag{13}$$

and define the coefficients β as volume-autocorrelation predictability. We expect β to be negative, and if it becomes more negative (or equivalently, if the absolute value of β becomes larger), we think more sniping (and thus more fast trading) is likely happening.

[29] In the extremely rare case where no transaction happens in the next second after a high volume hits the market, this paper looks from time $s+1$ to $s+2$, time $s+2$ to $s+3$, and so on until there are at least 2 transactions within a 1 second window.

A high volume of order canceling is a distinctive feature of HFT activities. cancel-to-trade is the ratio of submitted but canceled orders to executed trades of a certain stock from a particular exchange on a given day. For example, the IEX cancel-to-trade ratio for stock AAPL on day t is:

$$\text{cancel-to-trade} = \frac{\$ \text{ amount of AAPL orders submitted to IEX but canceled on day } t}{\$ \text{ amount of all AAPL orders executed by IEX on day } t} \tag{14}$$

Fundamental, slow traders buy and hold. HFT traders rarely hold a position for a very long time, and would typically unload within seconds. Without labeling of trader IDs, we wouldn't be able to track positions from different traders. However, we infer trade directions from trade and quote data using methods proposed by Lee & Ready (1991). For all the trading volume, we assume that orders reverse within 1 second are initiated by fast traders, and that the rest volume comes from slow traders[30]. The fundamental ratio is defined as the daily percentage of slow traders' volume divided by all volume. For example, the fundamental ratio of stock AAPL on day t is:

$$\text{Fundamental ratio} = 1 - \frac{\$ \text{ trading volume of AAPL that reverses within 1 second in IEX on day } t}{\$ \text{ total trading volume of AAPL in IEX on day } t} \tag{15}$$

4.1.4 Trading activity

Market share is a measure of how much share a particular exchange has in terms of trading a certain stock on a given trading day. It is roughly the popularity of an exchange. For example, the IEX market share for stock AAPL on day t is:

$$\text{Market Share of IEX} = \frac{\$ \text{ amount of IEX trading of AAPL on day } t}{\$ \text{ amount of trading of AAPL on day } t} \tag{16}$$

We use total trading volume as the proxy for how actively the market participants trade.

[30]This paper does not double count number of shares traded, so the "original" transaction must not be the "reversal" transaction of a previous "original" transaction. By definition, the "reversal" number of shares will never exceed the "original" number of shares traded, so the fast trader ratio always ranges between 0 and 1

4.2 Cross-sectional variations of stocks

Certain characteristics affect the price discovery and liquidity performance of a stock. For example, large market-cap, actively traded stocks tend to be more liquid, with faster price discovery (Chordia, Roll & Subrahmanyam (2000)). We control for these characteristics: market cap, volatility, and trading volume. Using data from CRSP, we construct a panel of each of the three measures for each stock/ trading day, starting 10 trading days before each event till 10 trading days after each event. We use a trailing average with a 10-trading-day window to calculate daily characteristics for each stock.

5 The impact of IEX speed bump

The first subsection studies the unconditional impact of the IEX speed bump on all stocks. Although with only moderate statistical significance, the results agrees with our hypothesis. The next subsection uses a diff-in-diff regression to test the impact on stocks that are heavily traded on IEX, and finds that almost all impacts are in the same direction but with much larger magnitudes. The following sections look at how cross-sectional variations for stocks affect the impact of the IEX speed bump and finds that the market quality for more volatile stocks and transactions happening closer to IEX or the SIP improves more after the IEX speed bump.

IEX has had a number of distinctive characteristics since its inception in 2012 as an Alternative Trading System (ATS). It does not offer co-location services, nor does it pay rebates to liquidity providers or liquidity takers. But most notably, IEX has implemented a speed bump since its inception. When IEX was included in the national market system starting Aug. 19th, 2016, the speed bump suddenly became relevant to all other public exchanges and market participants everywhere, since the quotes in IEX were required to be honored for Rule 611 ("no trade-through"). We investigate in this section the market reaction to this event. Notice the event is actually not the speed bump per se, but the inclusion of an exchange with a speed bump to the National Market System, which requires unified best price (NBBO) market-wise. For the sake of simplicity, we will still call it the impact of a speed bump in this paper.

5.1 Phase-in procedure

When IEX became a public exchange in 2016, it followed a phase-in procedure:[31]

- Fri, Aug 19: Two securities (VG, WIN)
- Wed, Aug 24: Eight securities (VALE,P,VHC,VIAV,VIP,VLY,XOMA,YINN,ZIOP)
- Mon, Aug 29: All other symbols that start with 'Y'-'Z'
- Wed, Aug 31: All other symbols that start with 'V'-'X'
- Fri, Sep 2: All other symbols that start with 'A'-'U'

Once a stock symbol is "phased in," IEX starts to report local best quotes to and receive NBBO from the public data server, the SIP. In addition, the "no trade-through" rule guarantees the execution price of market orders will be no worse than best bid and ask from IEX. Thus, the phase-in of IEX was the integration of IEX quotes into NBBO system. NBBO now depends on the price formation mechanism of IEX, which is affected by its speed bump. Through this channel, the impact of the speed bump disseminates to all other public exchanges in the market.

We list out the characteristics of the latter three groups of stocks that phased-in on different dates, and test whether the fourth and fifth groups of stocks (V-X, A-U) are statistically different from the third group (Y-Z).[32] Table 2 shows several key market quality measures for these three groups of stocks, both inside IEX (using IEX pegged mid-price from IEX TOPS) and from whole market (using NBBO mid-price from daily TAQ).

[31] source: https://iextrading.com/trading/alerts/2016/042/

[32] Technically, we should compare whether all five groups are statistically significant from each other, but since the sample sizes for the first two groups are too small, we ignore them.

Table 2: Stock ticker initials and characteristic average

Venue	IEX			Market		
Ticker initials	Y-Z	V-X	A-U	Y-Z	V-X	A-U
NBBO relevant (%)	23.6	22.1	23.4	100	100	100
Quoted Spread (bps)	56	49	53*	51	53	48
Effective spread (bps)	40	43	38	46	49	39*
Autocorrelation (1sec)	0.191	0.187	0.198	0.213	0.176*	0.181
Autocorrelation (60sec)	0.146	0.147	0.153	0.131	0.139	0.137
Cancel-to-Trade (%)	26.4	25.6	21.4*	34.2	28.5*	30.1

This table covers all equities traded on public exchanges throughout the phase-in period. Summary statistics are unconditional averages calculated based on data from all exchanges (the left three columns) and from the IEX only (the right three columns). The sample period covers 2 weeks (10 trading days) before the event and runs from Aug 3, 2016 to Aug 16, 2016. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively.

In most cases, stock ticker initials do not have structural impact on the measures we use for market quality, either inside IEX or in the market. What we observe here is that stocks starting with A-U are not statistically different from stocks starting with V-X or Y-Z. Thus, dividing stocks into these groups is indeed "random": the IEX speed bump is not selectively applied to stocks with certain characteristics over the phase-in period. Since the initial letter of the sticker of any stock is mostly not endogenously correlated with any other characteristics of the stock or the speed bump implementation, the phase-in procedure serves as a quasi-natural experiment to test the impact of IEX's speed bump.

5.2 Methodology

The main regression we use to test the impact of the IEX speed bump is:

$$y_{i,t} = \alpha_i + \beta Speedbump_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \tag{17}$$

where $y_{i,t}$ are various measures of market activity and market quality. α_i is the stock fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $X_{i,t}$ are a group of controls for a stock, for which we use end-of-day market cap, all-day

trading volume and all-day volatility. These are frequently used control variables for market liquidity and price efficiency. Controls will be calculated using 10-trading day average before the speed bump for each stock.

This event study setup works thanks to the randomization from the IEX speed bump phase-in procedure. We compare the market quality measures of every stock before and after the National Market System officially include the IEX speed bump on that stock through the investigation into the regression coefficient β.

Notice here and in later regressions, the LHS comes from intra-day high-frequency data, while the RHS comes from daily data. To resolve this inconsistency, a daily average of LHS is calculated for each day, each stock. Specifically:

1. Market share is summary of total dollar-amount trading, aggregated over each day for each stock
2. NBBO relevant and NBBO setting is time weighted
3. Cancel-to-Trade is volume weighted
4. Quoted spread is time weighted
5. Roll spread and effective spread is volume weighted
6. Autocorrelations of 1 second and 60 seconds are volume weighted
7. Information ratio is volume weighted

The processed data is a panel of all publicly traded stocks' (roughly 3500) daily market quality measure and descriptive statistic, 10 trading days before IEX became a public exchange to roughly 10 trading days after (a total of roughly 20 trading days). In section IV, the regression is first applied to both trades & quotes of IEX and that of the entire market, in order to show not only the impact of the IEX speed bump on its own trading, but the impact in the market as well. Then, cross-stock variation is studied to see if conditional impact of a speed bump is larger for stocks with certain characteristics.

5.3 Summary statistics

The trading landscape does not seem to have changed much since IEX introduced the speed bump, both inside and outside IEX. Although the market share of IEX has increased, it is mostly due to a time fixed effect which captures the trend that IEX gained publicity and grew steadily over August and September 2016, when IEX became more widely-known and accessible to investors as it became a public exchange. Since then, IEX market share has grown from 1% to 3% over five years (See Figure 2).

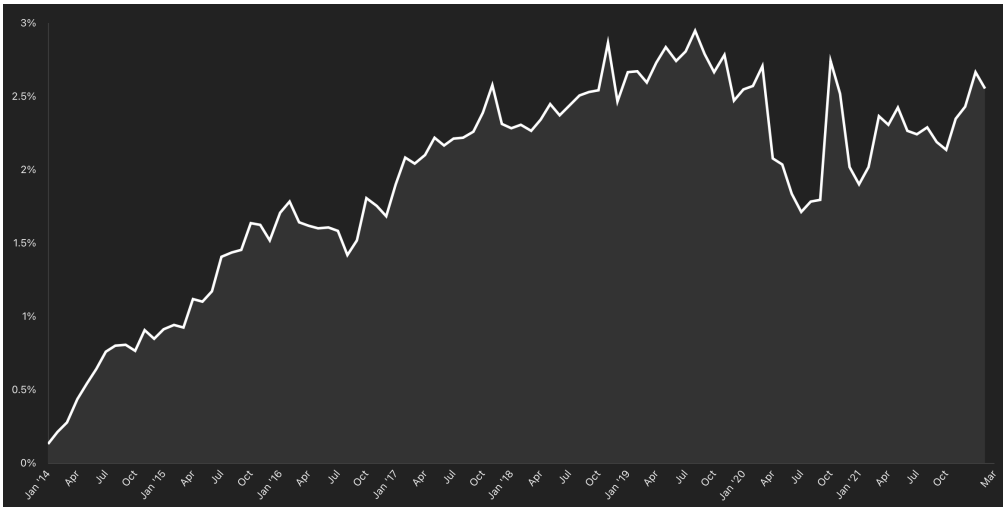


Figure 2: IEX market share for U.S. stocks Source: IEX

Table 3: Equity trading: before and after IEX became a public exchange

Characteristic averages	Before		After	
Venue	IEX	Market	IEX	Market
Quoted spread (bps)	61	57	59	58
Roll spread (bps)	37	38	40	35
Effective spread (bps)	42	39	41	40
Price impact (bps)	29	27	26	28
Realized spread (bps)	13	12	15	12
Volume-autocorrelation predictability	-1.26	-1.73	-0.98	-1.76
Cancel-to-Trade (%)	37.1	40.3	37.5	41.4
Fundamental ratio (%)	39.7	38.6	40.5	38.3
NBBO relevant (%)	23.6	100	26.1	100
NBBO setting (%)	6.4	100	6.6	100
Autocorrelation (1sec)	0.201	0.187	0.199	0.189
Autocorrelation (60sec)	0.146	0.153	0.142	0.165
Information ratio (%)	54.1	57.2	62.0	59.8
Market Share (%)	1.52	100	1.76	100

This table covers equities traded on IEX compared to all equities in the market throughout the phase-in period. The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 3, 2016 to Sept 16, 2016.

5.4 First difference: results

Based on our hypothesis, the IEX speed bump should improve liquidity and price discovery both in IEX and in the market, benefit slow traders, and attract more investors to IEX. Now we look at the evidence.

Table 4: Liquidity impact of IEX speed bump

Venue	IEX					Market				
	Quoted spread	Roll spread	Effective spread	Price impact	Realized spread	Quoted spread	Roll spread	Effective spread	Price impact	Realized spread
Speed bump	0.17	-0.02*	-0.27	-0.07*	-0.91	1.21	-0.05*	-0.82**	-0.29*	-1.14***
	(0.2)	(-1.8)	(-1.1)	(-1.6)	(-1.3)	(0.2)	(-1.6)	(-2.2)	(-1.9)	(-5.6)
Market cap	-2.21***	0.32**	0.12*	0.17	0.34	-14.2***	-12.5*	-4.3*	0.24	0.15
	(5.4)	(2.5)	(1.8)	(1.3)	(1.2)	(-9.2)	(-1.9)	(-1.6)	(1.5)	(0.7)
Volume	-0.43**	-0.25**	-1.53**	0.98	0.43	2.85	2.53**	-0.07	0.45	1.01
	(-2.2)	(-2.6)	(-2.1)	(1.5)	(0.8)	(1.3)	(2.3)	(-1.5)	(1.5)	(1.4)
Volatility	0.02**	0.11**	-0.002**	0.65	1.93	0.45*	1.21	-4.32**	0.26	0.96
	(1.9)	(2.2)	(-2.3)	(0.4)	(1.2)	(1.7)	(0.4)	(-2.5)	(1.5)	(1.4)
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.52	0.33	0.45	0.38	0.46	0.13	0.23	0.38	0.30	0.56

This table shows liquidity change when IEX, the first exchange with a speed bump, became a public exchange. We regress all measures of market activity and market quality on the speed bump implementation and characteristics of stocks:

$$y_{i,t} = \alpha_i + \beta Speedbump_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \quad (18)$$

where $y_{i,t}$ are variables in the first row of the table. α_i is the stock fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.
Quoted spread is the displayed bid-ask spread, Roll spread is implicit spread from trade price, while Effective spread is calculated from execution price against 60-second-before NBBO. All are liquidity indicators measured in percentage points.
The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 3, 2016 to Sept 16, 2016. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

There is no significant change in the quoted spread on IEX or the market. The stocks traded on IEX are biased towards bigger market-cap, liquid stocks. Most of those stocks' bid-ask spread is already the minimum one penny before IEX speed bump. This might be one reason that we do not see spread drop in the quoted spread. The Roll spread drops both inside (-0.02%) and outside (-0.05%) IEX, suggesting a slightly improved displayed liquidity. The effective spread changes very little inside IEX but decreases in the market (-0.82 %). One explanation of the different impact of spread change inside IEX could be that there is a trade-off of having a more accurate tracking of best price (thus tighter spread) over the market and having that price slower (thus a wide spread for a longer time period until updated) than an exchange without an artificial delay. In some cases one effect dominates the other. Price impact drops both inside (-0.07%) and outside (-0.29%) IEX, meaning that a speed bump would cause a less dramatic price change after an informational shock. Realized spread inside IEX does not change much, but decreases outside IEX (-1.14%), meaning that the market maker inside IEX is better off than outside market maker under the impact of a speed bump.

The paradoxical observation that effective spread improves while the quoted spread remains the same is driven by both the nature of National Market System and the design of the IEX speed bump. Although

the quoted spread could be different at different exchanges, Regulation NMS requires orders to be routed to the best price available anywhere within the National Market System. The IEX speed bump is designed to facilitate the access to best price (or equivalently, lowest effective spreads) by limiting high-frequency order sniping, but market makers at the IEX do not necessarily always give the best quotes.

Table 5: Price discovery impact of IEX speed bump

Venue	IEX			Market		
	Autocorr(1sec)	Autocorr(60sec)	Information ratio	Autocorr(1sec)	Autocorr(60sec)	Information ratio
Speed bump	-2.64***	-2.34	0.7	-1.24**	3.45*	0.34*
	(-4.3)	(-0.8)	(1.3)	(-2.5)	(1.7)	(1.9)
Market cap	-3.23	0.56*	2.34**	-16.2***	-31.2***	24.04*
	(-0.8)	(1.7)	(2.5)	(-3.2)	(-4.5)	(1.9)
Volume	4.42**	4.34	-1.03	6.78***	-4.23**	2.34*
	(2.1)	(1.5)	(-1.2)	(5.8)	(-2.4)	(1.8)
Volatility	-0.03*	0.05**	0.21**	-0.34**	1.4***	0.52
	(-1.9)	(2.1)	(2.4)	(-2.4)	(10.0)	(0.9)
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.62	0.25	0.19	0.66	0.24	0.20

This table shows price discovery change when IEX, the first exchange with a speed bump, became a public exchange. We regress all measures of market activity and market quality on the speed bump implementation and characteristics of stocks:

$$y_{i,t} = \alpha_i + \beta Speedbump_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \tag{19}$$

where $y_{i,t}$ are variables in the first row of the table. α_i is the stock fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.
Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.
The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 3, 2016 to Sept 16, 2016. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

In general, the IEX speed bump improves both price discovery and liquidity, not only on IEX itself, but on the entire market. Autocorrelation for prices every second drops both inside (-2.64%) and outside (-1.24%) IEX. Autocorrelation for prices every minute, however, increases in the market (3.45%). These might suggest that IEX speed bump improves short-lived price efficiency, while its impact on long-lived efficiency is not clear. Or it could be an indication that HFT are doing more arbitrage outside IEX to compensate for the part that they forgo due to the speed bump. Information ratio does not change significantly on IEX,

while increases a little in the market (0.34%), indicating slightly better price efficiency.

Table 6: Fast vs. slow trading impact of IEX speed bump

Venue	IEX			Market		
	Sniping	Cancel-to-Trade	Fundamental ratio	Sniping	Cancel-to-Trade	Fundamental ratio
Speed bump	0.174*	0.23	0.89	0.078***	2.23*	1.23**
	(1.8)	(0.8)	(1.3)	(3.0)	(1.7)	(2.2)
Market cap	5.20 *	1.42	2.12 *	-1.74	2.17	3.21
	(1.8)	(1.2)	(1.9)	(-1.1)	(1.5)	(1.1)
Volume	2.34	-1.76 **	2.95	4.75*	-2.85	3.65
	(1.1)	(-2.6)	(0.9)	(1.8)	(-1.5)	(1.4)
Volatility	0.04 **	-0.05***	0.22	-0.45	1.01	0.54 *
	(2.5)	(-2.9)	(1.4)	(-0.4)	(1.0)	(1.6)
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.54	0.32	0.41	0.34	0.17	0.18

This table shows fast vs. slow trading change when IEX, the first exchange with a speed bump, became a public exchange. We regress all measures of market activity and market quality on the speed bump implementation and characteristics of stocks:

$$y_{i,t} = \alpha_i + \beta Speedbump_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \tag{20}$$

where $y_{i,t}$ are variables in the first row of the table. α_i is the stock fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.
Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.
The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 3, 2016 to Sept 16, 2016. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

After the IEX speed bump, when a large volume happens, price autocorrelation is less negative both in IEX (0.174) and in the market (0.078), potentially due to lower level of HFT sniping. Cancel-to-trade ratio drops in the market (-2.23%) but not inside IEX. The fundamental ratio change in IEX is not statistically significant, but outside IEX it goes up by 1.23%. These suggest that the inclusion of the IEX speed bump discourages the quote sniping activities by fast traders in the aggregate market.

Table 7: Trading activity impact of IEX speed bump

Venue	IEX				Market	
	Share	NBBO relevant	NBBO setting	Cancel-to-Trade	Trading Volume	Cancel-to-Trade
Speed bump	0.023**	0.03*	0.07	-1.73**	1.47**	-0.68
	(2.4)	(1.7)	(0.8)	(-2.5)	(2.3)	(-1.0)
Market cap	-1.23***	0.17*	0.12*	-26.1***	-21.2***	-1.04**
	(-7.8)	(1.9)	(1.6)	(-9.2)	(-8.9)	(-2.1)
Volume	0.45***	0.23***	-0.73**	-1.73***	1.47	-0.3*
	(3.2)	(5.1)	(-2.3)	(-3.5)	(1.2)	(-1.7)
Volatility	-0.002***	0.001**	0.005***	-0.3	0.1	-1.25***
	(-2.9)	(2.5)	(2.7)	(-0.4)	(0.6)	(-3.5)
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.46	0.36	0.28	0.42	0.24	0.41

This table shows market activity and quality change when IEX, the first exchange with a speed bump, became a public exchange. We regress all measures of market activity and market quality on the speed bump implementation and characteristics of stocks:

$$y_{i,t} = \alpha_i + \beta Speedbump_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \tag{21}$$

where $y_{i,t}$ are variables in the first row of the table. α_i is the stock fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.
Share is the percentage of trading volume of IEX versus the market. NBBO relevant is the percentage of quote records of IEX that are inside or at NBBO. NBBO setting is the percentage of quote records of IEX that are better than previous NBBO and defines the new NBBO. Cancel-to-trade is the ratio of canceled orders versus trades. These trading activity indicators are all measured in % terms.
The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 3, 2016 to Sept 16, 2016. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

After the speed bump, the market share of IEX goes up by 0.023%, while total trading volume (in terms of total dollar amount) in the market goes up by 1.47% as well. Trading activity is not negatively affected by the artificial latency in IEX. During a day's trading, the quotes on IEX are only 0.03% more often to be able to match the NBBO, and IEX is still relatively unimportant in setting the NBBO price just as before. HFT activity, as indicated by cancel-to-trade ratio, drops both inside (-1.73%) and outside (-0.68%) IEX, although the change is only statistically significant inside IEX. The IEX speed bump does not disturb the market, except for discouraging HFT from it somewhat.

One pattern prevails in the previous three tests: the IEX speed bump is affecting the quotes from other

exchanges, sometimes even more significantly than the impact it has on itself. One possible explanation is phantom liquidity story: when there were no speed bump in the market, a investor seeking to trade a large amount won't be able to capture the full displayed liquidity. The first batch of trading would trigger latency arbitrage by HFT, which drives the price of the consequent trading worse for the investor. In anticipation for that, investors and market maker come up with an optimal quoting scheme. However, with IEX inside the public market, the protection is stronger, and thus might affect the way market markers quote or the way investors trade.

These results can be used to address some of the concerns in comment letters in response to the approval of IEX as a public exchange. One concern was the $350\mu s$ delay was too large to be "de minimus," that it could deter investors from trading. This is not the case in the sample period, where neither the market total trading volume nor the IEX market share declined. Another concern was the delay of information dissemination from IEX would have a bad impact on other exchanges benchmarking the NBBO. Although the change of quoted spread and Roll spread was not significant, our results for effective spread shows the liquidity improves both on IEX and on the entire market. The price efficiency does not change significantly on IEX itself, but improved on the entire market. There were also concerns that the speed bump would give a certain group of participants unfair advantage over others. We address this in a later section when we discuss the cross-sectional variations of the NYSE American speed bump.

5.5 Difference-in-difference: results

In the previous setup, some measures do not exhibit statistically significant change. The reason is that exchanges have very different market power over certain stocks. If a stock is thinly traded on IEX, we should not expect much of a difference the speed bump would make. Following this idea, we do a diff-in-diff regression to further explore the impact of a speed bump, especially on those stocks that has been heavily traded on IEX before it becomes public:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * MarketShare_{i,t} + \eta MarketShare_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \qquad (22)$$

where $y_{i,t}$ are variables of measures of market quality and activities. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $MarketShare_{i,t}$ is the average trading volume market share quantile of a stock on a particular day. $Speedbump_{i,t} * MarketShare_{i,t}$ is the interaction term that describes the diff-in-diff effect of the speed bump impact on high IEX market share stocks. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility. Controls will be calculated using 10-trading day average before the speed bump for each stock.

Again, we have a processed panel of stock market quality measures and characteristics for each stock on each trading day around the event of IEX became a public exchange. In addition to the first difference case, each stock now has a Market Share feature that is defined as the total dollar amount percentage traded on IEX versus the entire market over a rolling window of 10 trading days before a certain date. The regression coefficients thus tells us how actively traded stocks on IEX are affected by the IEX speed bump relative to other stocks.

Table 8: Diff-in-diff: The impact of IEX speed bump

Trading activity	IEX				Market	
	sniping	NBBO relevant	NBBO setting	Cancel-to-Trade	sniping	Cancel-to-Trade
Speed bump	0.07	1.24	0.32*	-1.54*	-1.65	-1.37
	(1.4)	(1.3)	(1.9)	(-1.8)	(-1.4)	(-1.3)
Speed bump ×	0.13***	0.18**	0.04	-1.43***	2.43*	-1.12***
Market Share	(4.6)	(2.5)	(1.5)	(-7.6)	(1.6)	(-3.5)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.43	0.33	0.52	0.27	0.39	0.27

Liquidity	IEX			Market		
	price impact	Roll spread	Effective spread	price impact	Roll spread	Effective spread
Speed bump	0.14	-1.31	-0.18	-1.62	-1.86	-1.49
	(3.6)	(-4.9)	(-2.7)	(-3.2)	(-4.7)	(-1.5)
Speed bump ×	0.17	-0.79***	-2.48***	-0.59*	-1.59*	-2.93***
Market Share	(1.2)	(-9.3)	(-12.4)	(-1.6)	(-1.7)	(-3.3)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.31	0.18	0.19	0.24	0.54	0.45

Efficiency	IEX			Market		
	Autocorr(1sec)	Autocorr(60sec)	Information ratio	Autocorr(1sec)	Autocorr(60sec)	Information ratio
Speed bump	-1.3***	0.001	0.001*	-0.3**	-1.21	0.81
	(-4.3)	(1.3)	(1.7)	(-2.1)	(-0.6)	(1.5)
Speed bump ×	-0.46***	-0.06*	2.15***	-0.38***	0.12	1.85**
Market Share	(-5.4)	(-1.7)	(4.3)	(-5.9)	(1.5)	(-2.5)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.27	0.16	0.47	0.25	0.24	0.33

This table shows market activity and quality change when IEX, the first exchange with a speed bump, became a public exchange. We regress all measures of market activity and market quality on the speed bump implementation, the market share of IEX of stocks, and characteristics of stocks:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * MarketShare_{i,t} + \gamma X_{i,t} + \eta MarketShare_{i,t} + \epsilon_{i,t} \quad (23)$$

where $y_{i,t}$ are variables in the first row of the table. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $MarketShare_{i,t}$ is the average trading volume market share quantile of a stock on a particular day. $Speedbump_{i,t} * MarketShare_{i,t}$ is the interaction term that describes the diff-in-diff effect of IEX market share on the speed bump impact. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.

IEX Share is the percentage of trading volume of IEX versus the market. NBBO relevant is the percentage of quote records of IEX that are inside or at NBBO. NBBO setting is the percentage of quote records of IEX that are better than previous NBBO and defines the new NBBO. Cancel-to-trade is the ratio of canceled orders versus trades. These trading activity indicators are all measured in % terms. Quoted spread is the displayed bid-ask spread, Roll spread is implicit spread from trade price, while Effective spread is calculated from execution price against 60-second-before NBBO. All are liquidity indicators measured in percentage points. Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.

The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 3, 2016 to Sept 16, 2016. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

Coefficients from diff-in-diff regression are more significant, the signs of which agree mostly with the first difference case. The market share of actively traded stocks on IEX increases 0.13% more than other stocks. Best quotes from IEX of actively traded stocks on IEX are 0.18% more often to match the NBBO. Even for stocks heavily traded on IEX, the exchange does not contribute much to the NBBO price setting. HFT activity drops more for IEX active stocks both inside (-1.43%) and outside (-1.12%) IEX.

The Roll spread drops more for IEX active stocks both inside (-0.79%) and outside (-1.59%) IEX, as does effective spread (-2.48% inside, -2.93% outside). In the aggregate market, the quoted spread of IEX active stocks drops more (-0.59%) than non-active stocks does. Inside the IEX, neither active nor non-active stocks' quoted spreads change much after the speed bump.

The Information ratio increases more for IEX active stocks both inside (2.15%) and outside (1.85%) IEX. The long-lived autocorrelation in the market does not change much for IEX active stocks; however, short-lived autocorrelation drops more for IEX active stocks both inside (-0.46%) and outside (-0.38%) IEX. Both results indicate better liquidity and efficiency after the implementation of the IEX speed bump.

The results show that impact of a speed bump acts similarly on stocks that are heavily traded on IEX before, but with a larger magnitude. This also adds to the evidence that a speed bump could potentially improve the market quality.

5.6 Cross-sectional variations: results

So far, our results show that the net effect of the IEX speed bump on market quality is positive. In this part, cross-sectional variations of stocks are utilized to explore the channels through which the speed bump has an impact trading activities. Specifically, information asymmetry and liquidity provision channels are tested to be both strengthening the overall impact in our results.

5.6.1 Volatility effects

In a theory model that introduces a speed bump on one of the two exchanges, Zhu (2019) predicts that stocks with higher volatility are more affected by the introduction of a speed bump. This is because a volatile stock has more information asymmetry before trading happens, and thus the magnitude of price discovery is in general higher in trading. The arbitrage activities by HFT levies a tax on the profitability of informed traders, deterring their willingness to reveal information. When a speed bump restrains the arbitrage ability of HFT by slowing them down, we would expect better improvement on price discovery and market efficiency for more volatile stocks. In addition, HFT quote snipers play a bigger role in volatile stocks, because higher volatility means more opportunities and larger magnitude of price discrepancies to arbitrage. Market makers require larger spread to compensate for more aggressive HFT arbitrage. As a mechanism designed to slow down the HFT, we should expect the speed bump to improve the liquidity more for a more volatile stock.

In order to test these hypothesis we run the regression:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * Volatility_{i,t} + \eta Volatility_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \tag{24}$$

where $y_{i,t}$ are variables of measures of market quality and activities. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $Volatility_{i,t}$ is the 10-trading-day moving average of stock daily return volatility of a stock on a particular day, normalized by the 10-trading-day average volatility of that stock pre-speed bump. $Speedbump_{i,t} * Volatility_{i,t}$ is the interaction term that describes the diff-in-diff effect of the speed bump impact on stocks with more informational contents (more price uncertainty). $X_{i,t}$ are a group of controls for a stock, for which we use market cap and volume. Controls will be calculated using 10-trading day average before the speed bump for each stock.

The diff-in-diff panel is very similar to the previous setup, except that we substitute the additional feature of IEX active trading with ex-ante volatility (over a rolling window of 10 trading days before a certain date) as an proxy for information content variation of a certain stock. The regression coefficients

thus tells us how are the more volatile stocks affected by the IEX speed bump relative to other stocks. Results are shown in Table 9.

Table 9: Volatility variations: The impact of IEX speed bump

Trading activity	IEX				Market	
	sniping	NBBO relevant	NBBO setting	Cancel-to-Trade	sniping	Cancel-to-Trade
Speed bump	1.23	0.52*	1.45	-4.24*	-2.53	-0.97**
	(1.5)	(1.8)	(0.6)	(-1.9)	(-0.8)	(-2.5)
Speed bump ×	0.25**	0.33	0.42***	0.74	-2.54*	-1.54*
Effective Volatility	(2.6)	(1.5)	(15.7)	(0.8)	(-1.6)	(-1.7)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.26	0.43	0.52	0.29	0.16	0.32

Liquidity	IEX			Market		
	price impact	Roll spread	Effective spread	price impact	Roll spread	Effective spread
Speed bump	-2.4*	1.37	1.22	-1.65	-1.12**	-2.05
	(-1.8)	(0.9)	(1.2)	(-1.2)	(-2.2)	(-1.1)
Speed bump ×	-0.45***	0.23***	-0.73***	-1.73***	1.47	-0.3**
Effective Volatility	(-3.2)	(6.1)	(-9.2)	(-3.5)	(1.2)	(-2.3)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.19	0.46	0.42	0.35	0.40	0.29

Efficiency	IEX			Market		
	Autocorr(1sec)	Autocorr(60sec)	Information ratio	Autocorr(1sec)	Autocorr(60sec)	Information ratio
Speed bump	-0.02	0.13*	2.21	-0.56	0.23**	-1.64
	(-1.4)	(1.8)	(1.5)	(-0.5)	(2.3)	(-1.4)
Speed bump ×	-0.07***	-1.03**	-0.21**	-0.42***	-0.1	3.47***
Effective Volatility	(-12.5)	(-2.4)	(-2.2)	(-9.1)	(-1.5)	(2.9)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.45	0.24	0.16	0.37	0.56	0.26

This table shows market activity and quality change when IEX, the first exchange with a speed bump, became a public exchange. We regress all measures of market activity and market quality on the speed bump implementation, market share of IEX of stocks, and characteristics of stocks:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * Volatility_{i,t} + \gamma X_{i,t} + \eta Volatility_{i,t} + \epsilon_{i,t} \quad (25)$$

where $y_{i,t}$ are variables in the first row of the table. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $Volatility_{i,t}$ is the 10-trading-day moving average of stock daily return volatility of a stock on a particular day, normalized by the 10-trading-day average volatility of that stock pre-speed bump. $Speedbump_{i,t} * Volatility_{i,t}$ is the interaction term that describes the diff-in-diff effect of the speed bump impact on stocks with more informational contents (more price uncertainty). $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.

IEX Share is the percentage of trading volume of IEX versus the market. NBBO relevant is the percentage of quote records of IEX that are inside or at NBBO. NBBO setting is the percentage of quote records of IEX that are better than previous NBBO and defines the new NBBO. Cancel-to-trade is the ratio of canceled orders versus trades. These trading activity indicators are all measured in % terms. Quoted spread is the displayed bid-ask spread, Roll spread is implicit spread from trade price, while Effective spread is calculated from execution price against 60-second-before NBBO. All are liquidity indicators measured in percentage points. Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.

The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 3, 2016 to Sept 16, 2016. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

Volatile stocks' total trading volume in the market is 2.54% less than before the speed bump, while IEX trades 0.25% more of volatile stocks after the implementation of the speed bump. IEXs ability to match NBBO for volatile stocks does not change significantly, while it contributes 0.42% more to the NBBO price setting for those stocks. These two observations suggest IEX plays a more important role in volatile stock trading after the speed bump. The cancel-to-trade ratio of volatile stocks does not exhibit significant change, perhaps because volatile stocks are inherently favored by HFT even in the presence of the delay of IEX speed bump. The quoted spread drops more for volatile stocks both inside (-0.45%) and outside (-1.73%) IEX. The effective spread drop more for volatile stocks both inside (-0.73%) and outside (-0.3%) IEX as well. Roll spread increases a bit for volatile stocks on both IEX (0.23%) and the market (1.47%). This could be because trading prices of volatile stocks oscillates between bid and ask and thus incur higher implicit cost. Autocorrelation for price every second (-0.07%) and every minute (-0.42%) drops more for volatile stocks after the speed bump inside IEX. Information ratio for volatile stocks increases on IEX (0.21%) but decreases on the entire market (-3.47%). In general, liquidity and price efficiency improve more for stocks with higher volatility.

5.6.2 Location effects

The servers of different exchanges are located in different places. The time it takes for a piece of information to be transmitted from the SIP to an exchange differs. Ernst, Sokobin & Spatt (2022) exploit the structure of geographic latencies to pinpoint the price discovery patterns. Similarly, this paper uses geographic variation of different exchange servers to further unravel the speed bump effects on market quality.

In the daily TAQ data, each entry has two timestamps: one that marks the broadcasting time of a signal, and one that marks the time that signal gets to destination. To estimate the distances to the SIP (transmission time to or from the SIP) of all public exchanges, we take the average of the differences between two timestamps for all trades and quotes on Aug. 26, 2016, which is summarized in Figure 3.

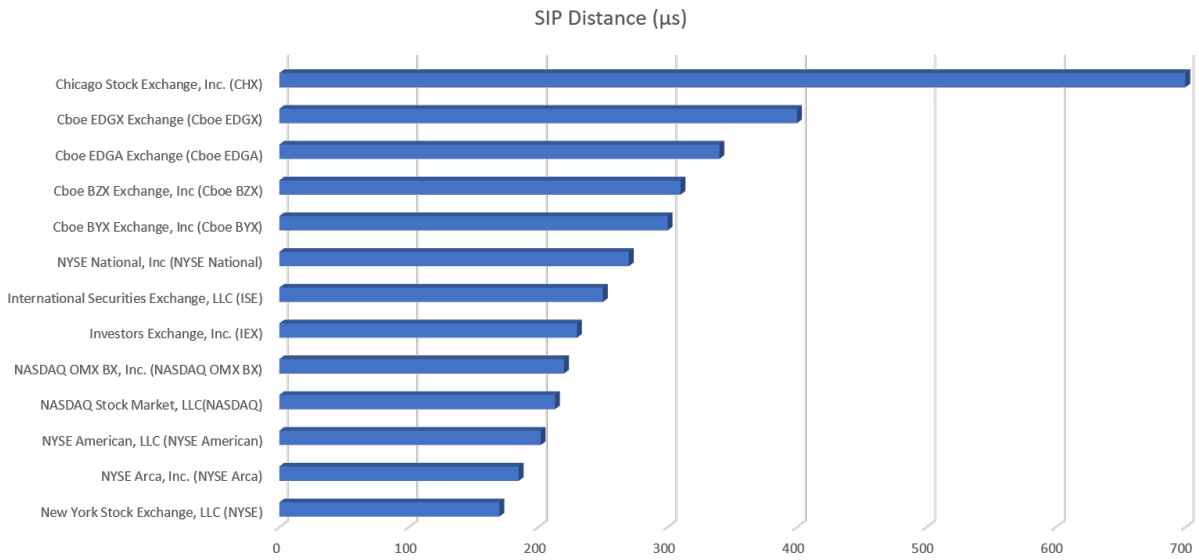


Figure 3: Distance to the SIP for public equity exchanges: estimation from daily TAQ timestamps

We predict the distance to the SIP to be inversely related to the magnitude of the impact of a speed bump. The speed bump of IEX encourages price discovery, which contributes to a more accurate NBBO at the very time IEX updates its quotes to the SIP. If an exchange is very close to the SIP, the updated, more accurate information might be of very high significance. But for an exchange that is far away, the accurateness of NBBO on the SIP might matter less, because price might fluctuate again or the HFT might front-run that long distance.

A liquidity provider would have more assurance when faced with less severe adverse selection problem. In other words, one is willing to provide cheaper liquidity or larger amount of liquidity if the quote is less likely to be stale and vulnerable to HFT sniping. Thus, we predict market efficiency and liquidity on exchanges closer to the SIP to have larger improvement after the IEX speed bump.

An alternative perspective is the relative distances among different exchanges. Would the IEX speed bump have the same impact on an exchange that's located very close to IEX vs. an exchange that's very far away? Zhu (2019) predicts that if two exchanges are closer to each other, the impact of a speed bump is larger. Quotes on two close exchanges are vulnerable to more aggressive HFT arbitrage, since HFT can

quickly unload position and only have to keep inventory for a very short amount of time. When a speed bump is introduced, we would expect the protection to encourage liquidity provision and price discovery which contribute to the improvement of market efficiency and liquidity.

The geographical locations of U.S. equity exchanges is shown in Figure 4:[33]

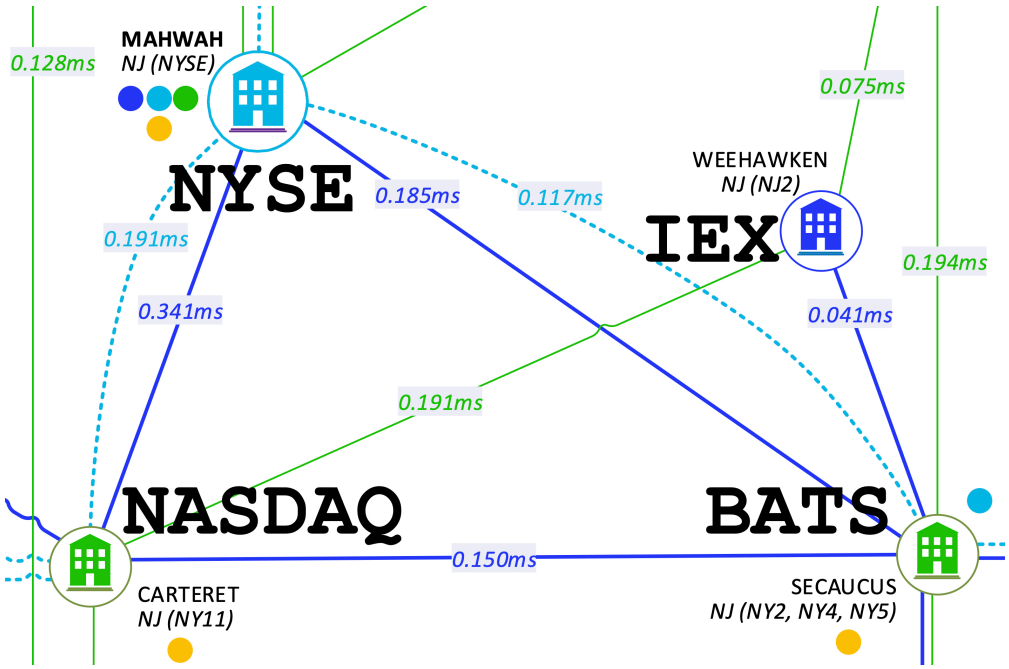


Figure 4: Geographical topology of U.S. equity exchanges
Mahwah, Weehawken, Carteret and Secuaucs are four locations in interest. NJ2, NY2, NY4, NY5, NY11 are code names for data centers. Each exchange family occupy one data center.

Table 10 summarizes the relative distances and signal transmission latencies between each two locations:

[33]source: https://www.theice.com/market-data/connectivity-and-feeds/network-topology-map. Some exchanges offer faster access outside their primary data centers (Point of Presence, or POP) that allow customers to connect to different exchanges with lower latency. For example, IEX POP is located in data center NY5, Secaucus; Nasdaq POP is located in data center NY4, Secaucus. For simplicity we only list out the location for the matching engine of each exchange.

Table 10: Geographical distances and message travel time among major U.S. equity exchanges

Distances (km)	Mahwah	Secaucus	Weehawken	Carteret
Mahwah (NYSE)				
Secaucus (BATS)	33.44			
Weehawke (IEX)	36.72	4.44		
Carteret (NASDAQ)	54.77	25.65	26.82	

Light-speed travel time (μs)	NYSE	BATS	IEX	NASDAQ
NYSE				
BATS	111.5			
IEX	122.4	14.8		
NASDAQ	182.6	85.5	89.4	

Distance measured in Google Map. Travel time calculated by dividing distance with the speed of light.

Exchanges in Chicago are too far away from the rest of the market and only have less than 0.1% market share of equity trading, so we will ignore those exchanges for this part. According to the locations of their data servers, exchanges are grouped into 3 categories (Mahwah, Secaucus, and Carteret), in terms of how far they are from IEX. We define it as the relative distance to the IEX, which provides an alternative description of the geographical locations for different exchanges.

In order to test the hypotheses that:

1. The distance to the SIP is inversely related to the magnitude of the impact of IEX speed bump
2. Relative distance to the IEX is inversely related to the magnitude of the impact of IEX speed bump

We run the regression:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * Distance_{i,t} + \eta Distance_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \tag{26}$$

where $y_{i,t}$ are variables of measures of market quality and activities. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for

all the days in the future). $Distance_{i,t}$ is measured by the Google earth geographical distance between two locations (whether that's the distance from the exchange that posts a quote/executes a transaction to IEX or its distance to the public SIP) divided by speed of light. $Speedbump_{i,t} * Distance_{i,t}$ is the interaction term that describes the diff-in-diff effect of the speed bump impact on trades that travel a long distance. $X_{i,t}$ are a group of controls for a stock, for which we use market cap and volume. Controls will be calculated using 10-trading day average before the speed bump for each stock.

The data structure here is different from previous setups. Here, we also track the *Distance*, determined by where (in the four main exchange server locations) the trade or quote happens. We record 1) how far the exchange executing or quoting is away from the SIP 2) how far the exchange executing or quoting is away from the IEX. *Distance 1* is measured by the average travel time of all transactions from 4 server locations to the SIP; *Distance 2* is measured by dividing the straight line distance on Google Earth from 4 server locations to each other with the speed of light. Both *Distance 1* and *Distance 2* are in microseconds (μs) units.

Previous regressions use a panel of (stock,day), where each data point is the daily average of the market quality measure of a particular stock. Here the regression is based on a panel of (stock-server location,day), where each data point is the daily and location average of the market quality measure of a particular stock.

Thus, the regression coefficients tells us differences in market performance change for exchanges in different locations after the IEX speed bump. Regression based on exchanges distance to the SIP and their relative distance to IEX is summarized in Table 11.

Table 11: Geographical location variations: The impact of IEX speed bump

Trading activity	distance to the SIP			Relative distance to IEX		
	snipng	fundamental ratio	Cancel-to-Trade	sniping	fundamental ratio	Cancel-to-Trade
Speed bump	1.56	1.36*	5.72*	-2.44	-3.45*	-3.52**
	(1.4)	(1.8)	(1.6)	(-0.8)	(-1.7)	(-2.5)
Speed bump ×	1.33**	0.25***	-2.43**	3.554**	1.87**	1.09***
Distance	(2.3)	(7.6)	(-2.4)	(2.5)	(2.5)	(4.4)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.43	0.24	0.28	0.32	0.56	0.48

Liquidity	distance to the SIP			Relative distance to IEX		
	price impact	Roll spread	Effective spread	price impact	Roll spread	Effective spread
Speed bump	-4.32	2.34	3.14	-1.19**	-3.12***	-4.01*
	(-0.8)	(0.9)	(1.2)	(-2.4)	(-5.2)	(-1.8)
Speed bump ×	-1.75***	2.48	-4.47**	-0.84***	-2.65*	-1.13**
Distance	(-4.2)	(1.1)	(-2.5)	(-8.8)	(-1.9)	(-2.2)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.42	0.27	0.47	0.35	0.33	0.22

Efficiency	distance to the SIP			Relative distance to IEX		
	Autocorr(1sec)	Autocorr(60sec)	Information ratio	Autocorr(1sec)	Autocorr(60sec)	Information ratio
Speed bump	-0.51**	0.22	2.32	-0.98*	1.14	4.64
	(-2.4)	(1.3)	(1.4)	(-1.9)	(1.5)	(0.7)
Speed bump ×	-2.21**	-3.45*	2.45***	0.45**	2.31	5.47***
Distance	(-2.4)	(-1.6)	(5.2)	(-2.6)	(1.5)	(4.7)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.19	0.25	0.16	0.39	0.28	0.32

This table shows market activity and quality change when IEX, the first exchange with a speed bump, became a public exchange. We regress all measures of market activity and market quality on the speed bump implementation, information transmission distance from exchanges to the SIP/IEX, and characteristics of stocks:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * D_{i,t} + \gamma X_{i,t} + \eta D_{i,t} + \epsilon_{i,t} \tag{27}$$

where $y_{i,t}$ are variables in the first row of the table. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $D_{i,t}$ describes the geographical location of an exchange relaitve to the SIP or IEX. $Speedbump_{i,t} * D_{i,t}$ is the interaction term that describes the diff-in-diff effect of geographical location difference on the speed bump impact. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.

NBBO relevant is the percentage of quote records of IEX that are inside or at NBBO. NBBO setting is the percentage of quote records of IEX that are better than previous NBBO and defines the new NBBO. Cancel-to-trade is the ratio of canceled orders versus trades. These trading activity indicators are all measured in % terms. Quoted spread is the displayed bid-ask spread, Roll spread is implicit spread from trade price, while Effective spread is calculated from execution price against 60-second-before NBBO. All are liquidity indicators measured in percentage points. Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.

The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 3, 2016 to Sept 16, 2016. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

After the implementation of a speed bump, an exchange closer to the SIP is more likely to quote as good or even improve the NBBO price. This agrees with the channel of liquidity provision that we discussed, since the closer exchanges enjoy better liquidity provision improvements. Quoted spread and Effective spread drops more, while Roll spread change on a closer exchange is not significantly different from a distant exchange. Both second and minute autocorrelation tends to drop more on an exchange closer to the SIP, although information ratio improves less. Apparently, the exchanges which are very close to the SIP enjoys larger market quality improvement from the IEX speed bump, while exchanges far away are less affected. In other words, as the benefit of obtaining accurate NBBO radiates from the center of the SIP, it slowly fades.

For exchanges close to IEX, the speed bump seems to discourage quoting and encouraging HFT activity. Spreads on exchanges near IEX drops more with the implementation of the speed bump. Autocorrelation is higher while information ratio is lower. One explanation for these results is that since the speed bump makes it harder to arbitrage on IEX, HFT tend to find substitutes for IEX and arbitrage more on nearby exchanges. More HFT arbitrage masks the price discovery, making the volatility noisier.

One caveat for the results: IEX happens to be located not far from the SIP. NYSE and Nasdaq are also located close to the SIP. Since those two big exchange groups dominate the trading volume, our results could potentially be driven by very large exchanges. However, the regression results are still informative in that they reveal the impact of a speed bump on the most important players in U.S. equity trading.

6 The impact of NYSE American speed bump

NYSE American specializes in trading of small to medium market cap stocks. It has over 370 listings of small-medium market cap stocks, and has a so-called Designated Market Maker (DMM) system, the members of which are specialists that commit to help facilitate the trading for certain stocks. In July 2017, NYSE American's speed bump went in effect, approximately one year after IEX's speed bump. Since this paper studies these two events by looking at roughly 10 trading days before and after the implementations of these speed bumps, there is no overlapping sample periods of these two events, and we assume there is no correlation

between them that would alter the interpretation of results of either regression. Although seemingly identical to the existing IEX speed bump, our test shows that these two speed bumps are significantly different.

6.1 Methodology

We use a diff-in-diff regression to investigate the impact of NYSE American's speed bump:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * MarketShare_{i,t} + \gamma X_{i,t} + \eta MarketShare_{i,t} + \epsilon_{i,t} \quad (28)$$

where $y_{i,t}$ are market activity and market quality measures. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has NYSE American speed bump implemented on a particular day (and for all the days in the future). $MarketShare_{i,t}$ is the average trading volume NYSE American market share quantile of a stock on a particular day. $Speedbump_{i,t} * MarketShare_{i,t}$ is the interaction term that describes the diff-in-diff effect of speed bump impact on high NYSE American market share stocks. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility. Controls will be calculated using 10-trading day average before the NYSE American speed bump for each stock.

Similarly to the case of IEX, we construct the diff-in-diff structure so that the regression would focus on stocks already heavily traded on NYSE American. Both the impact on NYSE American and the entire market are investigated. The situation was slightly different though. First, there was no phase-in procedure. Second, NYSE American was migrated to Pillar on the same date that the speed bump got implemented (which is actually powered by the new trading platform technology on Pillar). The diff-in-diff helps disentangle these two effects: any stock, as long as it is active on any of the NYSE exchanges, should be affected by migrating to Pillar; however, only stocks that are active on NYSE American will be affected by the speed bump (because there should always be relevant quotes to refer to). The difference between the treatment and the control would then reflect the impact of a speed bump.

6.2 Difference-in-difference: results

The results for the NYSE American speed bump are very different from the results for IEX speed bump. We do not observe liquidity or efficiency improvements, nor do we see a competition happening between IEX and NYSE American despite the fact that their speed bump implementation are almost identical on paper.

Table 12: Diff-in-diff: The impact of NYSE American speed bump

Trading activity	NYSE American				Market	
	sniping	NBBO relevant	NBBO setting	Cancel-to-Trade	sniping	Cancel-to-Trade
Speed bump	5.13	4.35*	2.43	2.54*	5.43	1.54
	(0.4)	(1.7)	(1.5)	(1.7)	(0.7)	(1.4)
Speed bump ×	2.52	0.54	-1.14	-0.42	3.35*	2.12
Market share	(1.3)	(1.4)	(-0.7)	(-1.4)	(1.9)	(0.6)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.23	0.34	0.46	0.13	0.41	0.20

Liquidity	NYSE American			Market		
	price impact	Roll spread	Effective spread	price impact	Roll spread	Effective spread
Speed bump	-4.23**	1.35*	3.52	-6.12*	-3.41*	-2.04
	(-2.5)	(1.6)	(1.4)	(-1.9)	(-1.7)	(-0.9)
Speed bump ×	3.23	0.84*	-0.34**	-2.53	3.54	-1.32
Market share	(0.6)	(1.7)	(-2.1)	(-1.4)	(1.2)	(-0.8)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.19	0.28	0.46	0.31	0.26	0.25

Efficiency	NYSE American			Market		
	Autocorr(1sec)	Autocorr(60sec)	Information ratio	Autocorr(1sec)	Autocorr(60sec)	Information ratio
Speed bump	0.24	0.03	0.14	-2.34	1.45	-6.37***
	(1.4)	(0.7)	(1.5)	(-1.4)	(1.2)	(-7.9)
Speed bump ×	-0.31	0.12**	-1.63***	0.45	3.22	-1.35
Market share	(-0.7)	(2.6)	(-3.5)	(1.4)	(0.6)	(-1.0)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.36	0.28	0.26	0.41	0.13	0.38

This table shows market activity and quality change when NYSE American introduced its speed bump. We regress all measures of market activity and market quality on the speed bump implementation, market share of NYSE American of stocks, and characteristics of stocks:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * MarketShare_{i,t} + \gamma X_{i,t} + \eta MarketShare_{i,t} + \epsilon_{i,t} \tag{29}$$

where $y_{i,t}$ are variables in the first row of the table. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). $MarketShare_{i,t}$ is the average trading volume market share quantile of a stock on a particular day. $Speedbump_{i,t} * MarketShare_{i,t}$ is the interaction term that describes the diff-in-diff effect of NYSE American market share on the speed bump impact. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.

Share is the percentage of trading volume of NYSE American versus the market. NBBO relevant is the percentage of quote records of NYSE American that are inside or at NBBO. NBBO setting is the percentage of quote records of NYSE American that are better than previous NBBO and defines the new NBBO.Cancel-to-trade is the ratio of canceled orders versus trades. These market activity indicators are all measured in % terms. Quoted spread is the displayed bid-ask spread, Roll spread is implicit spread from trade price, while Effective spread is calculated from execution price against 60-second-before NBBO. All are liquidity indicators measured in percentage points. Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.

The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from July 10, 2017 to August 10, 2017. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

The NYSE American does not cause statistically significant change in most market performance measures. NYSE Americans market share does not increase significantly. Effective spread drops, while Quoted and Roll spread stay the same. Autocorrelation increases while information ratio decrease. None of these changes are significant in the market.

One might have identification concern that the date of the speed bump was endogenously chosen by NYSE American to facilitate the trading. This was not the case according to the documents from NYSE. NYSE set up the road map of migrating its exchanges one by one to the new trading platform Pillar long before the NYSE American speed bump. It was unlikely that NYSE could predict NYSE American's trading far in the future on a specific day and introduced a speed bump based on that information.

As a robustness check, we hypothesize that there was a trend happening around the implementation of the NYSE American speed bump that shifts the market activity and/or market quality of stocks trading on NYSE American.

Table 13: NYSE American: placebo test

Coefficient on time fixed effect	All stocks	NYSE American active stocks
Market Share (%)	0.07*	0.07
NBBO relevant (%)	1.03	1.33
Quoted Spread (bps)	97	99
Effective spread (bps)	77	79
Autocorrelation (60sec)	0.03	0.09
Autocorrelation (60sec)	0.05	0.17
Cancel-to-Trade (%)	0.19**	-0.14

This table covers all publicly traded equities around NYSE American speed bump implementation, July 2017. NYSE American active stocks are defined as stocks with trading volume on top 1/3. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively.

From this table we see that most variables in interest has statistically insignificant change in time over the sample period. This would add assurance to the assumption that the NYSE American speed bump was an exogenous shock to the stock trading.

6.3 Comparison to IEX speed bump

When IEX was seeking approval of being a national exchange from the SEC, NYSE expressed strong objection.[34] NYSE vehemently opposed IEXs exchange application, saying the model was bad for the market, but later said it planned to copy aspects of the newest U.S. exchange in order to better compete.

For convenience, Table 14 summarizes empirical results from previous sections and compares the impact of the IEX speed bump and the impact of the NYSE American speed bump:

Table 14: IEX vs. NYSE American: Comparing two speed bumps

Event	IEX speed bump		NYSE American speed bump	
Venue	IEX	Market	NYSE American	Market
sniping	0.13***	N/A	2.52	N/A
NBBO relevant	0.18**	N/A	0.54	N/A
NBBO setting	0.04	N/A	-1.14	N/A
Cancel-to-Trade	-1.43***	-1.12***	-0.42	2.12
price impact	0.17	-0.59*	3.23	-2.53
Roll spread	-0.79***	-1.59*	0.84*	3.54
Effective spread	-2.48***	-2.93***	-0.34**	-1.32
Autocorr(1sec)	-0.46***	-0.38***	-0.31	0.45
Autocorr(60sec)	-0.06*	0.12	0.12**	3.22
Information ratio	2.15***	1.85**	-1.63***	-1.35

This table compares the diff-in-diff coefficients from the regressions of market activities and market quality changes on IEX and NYSE American speed bump implementation. It covers all publicly traded equities around IEX speed bump implementation, Aug-Sept 2016, and all publicly traded equities around NYSE American speed bump implementation, July 2017. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively.

We see that the speed bump of NYSE American does not improve the market quality. The relative market share of NYSE American and IEX remained the same after the introduction of NYSE American speed bump, which means that the speed bump is not helping NYSE American attract a specific group of investors that favor a market with speed bump.

One possible explanation is that NYSE American's speed bump, due to the existence of co-location, does not serve properly as an equal slowdown to everyone. The IEX does not offer co-location service, and thus

[34] https://www.reuters.com/article/us-nyse-exchange-speedbump/nyse-wins-regulatory-approval-for-speed-bump-exchange-idU.S.KCN18C2UC

its speed bump only allows the update of NBBO to bypass. Meanwhile, any investor who has subscription to co-location service of NYSE American will not be subject to this artificial delay. Since the HFT are most likely to be co-location subscribers, NYSE American's speed bump might not slow them down.

Another possibility is that the implementation with same amount of artificial delay ($350\mu s$) does not work with NYSE American because its matching engine is at different geographical location from IEX. IEX's matching engine is located in Weehawken, while NYSE American's matching engine is located in Mahwah. IEX chose $350\mu s$ because it is roughly the longest time it takes for a signal to travel from any other exchange to IEX and back. This delay gives IEX the time to react to quote changes from anywhere in the market. NYSE American uses the same coefficients and copied the IEX setup. If some of the setup only fits IEX but not NYSE American, it could cause the NYSE American to be not as effective.

We should not forget that the speed bump is only part of IEX's unique trading regime. On top of the $350\mu s$ artificial delay, IEX devises a signal that prevents its orders from being executed when a predatory trading pattern is detected.[35] The pegged order type is equipped with this technology and lessens the adverse selection problem of stale-quote sniping faced by NBBO-tracking investors. In December 2019, IEX filed another proposal for a new order type called "D-Limit", which aims to extend similar protection to displayed order.[36] None of these are adopted by NYSE American when it introduced its speed bump.

In addition, IEX does not pay any rebates while NYSE American follows a maker-taker rebate regime. Spatt (2019) argues the tiered rebate structure could potentially distort incentives of brokers. Wah & Feldman (2018) find empirical evidence that maker-taker pricing model can lead to long queues. Since fee structure and rebate regime may affect a broker's order routing decision, it might mask the impact of a speed bump. The routing data is proprietary to brokers and thus not publicly available. One reason for brokers' unwillingness to disclose routing data is because they are required by law to act at the best interest of their clients. Being exposed would draw more attention to their routing decision making. Besides, how a broker routes its orders is part of its business secret of executing its clients' orders in a most efficient way. Disclosing the routing data might give away the algorithm that can be costly to develop.

[35] see more in IEX white paper https://iextrading.com/docs/The Evolution of the Crumbling Quote Signal.pdf
[36] https://iextrading.com/docs/rule-filings/SR-IEX-2019-15.pdf

Finally, NYSE has a Designated Market Maker system while market making in IEX are completely decentralized. Specialist play an important role in providing liquidity on NYSE American, while more than half of IEX's trading volume comes from mid-price market orders. A speed bump could interact with the market making in different ways because of this distinction. These are all plausible but impossible to test without access to proprietary data. In next section, however, we propose one rationale that could motivate NSYE American to implement a speed bump, even if it does not help improve liquidity or efficiency.

6.4 Cross-sectional variations: results

One distinct feature of NYSE American, compared to IEX, is the Designated Market Maker system. Specialists registered in this system are obliged to provide liquidity for certain stocks, and thus are susceptible to stale quote sniping by HFT. With the help of a speed bump, market makers can update their view of the market, NBBO, before executing orders. If a specialist considers her quote to be stale, she will have more time to retract that order and put up a new one thanks to the speed bump. The artificial delay gives market makers a last-look privilege by alleviating the asymmetric information issue they face.

Thus, we should expect DMMs in NYSE American to enjoy better improvement from the introduction of a speed bump. Specifically, trades handled by specialists on NYSE American should improve more than other trades. If the hypothesis is true, market quality of specialist handled trades on NYSE American should improve more than other trades. Moreover, market quality of specialist handled trades should not change much outside NYSE American.

To test this, we regress all measures of market activity and market quality on the speed bump implementation, NYSE American DMM specialist handling dummy variable, and characteristics of stocks:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * DMM_i + \gamma X_{i,t} + \eta DMM_i + \epsilon_{i,t} \tag{30}$$

where $y_{i,t}$ are variables in the first row of the table. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days

in the future). DMM_i is the dummy variable that equals 1 if a stock is facilitated by NYSE American DMMs. $Speedbump_{i,t} * DMM_i$ is the interaction term that describes the diff-in-diff effect of whether a stock is handled by NYSE American DMMs on the speed bump impact. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility.

Table 15 shows the results from the previous regression:

Table 15: Specialists vs. decentralized: The impact of NYSE American speed bump

Trading activity	NYSE American				Market	
	sniping	NBBO relevant	NBBO setting	Cancel-to-Trade	sniping	Cancel-to-Trade
Speed bump	1.32	4.13*	3.41	3.41*	-4.32*	-1.65
	(1.5)	(1.7)	(0.7)	(1.7)	(-1.6)	(-1.5)
Speed bump ×	0.14***	0.31***	3.42	-2.56	-1.02	-5.24
Specialist dealing	(7.4)	(3.8)	(1.4)	(-0.8)	(-0.7)	(-1.1)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.54	0.17	0.32	0.14	0.24	0.17

Liquidity	NYSE American			Market		
	price impact	Roll spread	Effective spread	price impact	Roll spread	Effective spread
Speed bump	-1.43*	5.67*	0.12***	-2.12*	-0.32	-2.74*
	(-1.6)	(1.9)	(5.2)	(-1.8)	(-1.4)	(-1.8)
Speed bump ×	-1.27*	-0.42***	-0.64**	-1.03	0.10	-0.38
Specialist dealing	(-1.6)	(-4.0)	(-2.6)	(-1.5)	(0.9)	(-1.1)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.13	0.25	0.34	0.42	0.25	0.29

Efficiency	NYSE American			Market		
	Autocorr(1sec)	Autocorr(60sec)	Information ratio	Autocorr(1sec)	Autocorr(60sec)	Information ratio
Speed bump	0.42*	1.42	4.24	1.32**	1.65	-2.45
	(1.6)	(0.7)	(1.4)	(-2.5)	(0.6)	(-1.4)
Speed bump ×	-3.49***	1.61*	3.55**	-2.30	0.47	-1.14
Specialist dealing	(-4.6)	(1.8)	(2.5)	(-0.6)	(1.0)	(-0.4)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.24	0.17	0.26	0.47	0.35	0.21

This table shows market activity and quality change when NYSE American debuted its speed bump. We regress all measures of market activity and market quality on the speed bump implementation, market share of NYSE American of stocks, and characteristics of stocks:

$$y_{i,t} = \alpha_t + \delta Speedbump_{i,t} + \beta Speedbump_{i,t} * DMM_i + \gamma X_{i,t} + \eta DMM_i + \epsilon_{i,t} \quad (31)$$

where $y_{i,t}$ are variables in the first row of the table. α_t is the time (trading day) fixed effect. $Speedbump_{i,t}$ is a binary variable indicating whether a stock has a speed bump on a particular day (and for all the days in the future). DMM_i is the dummy variable that equals 1 if a stock is facilitated by NYSE American DMMs. $Speedbump_{i,t} * DMM_i$ is the interaction term that describes the diff-in-diff effect of whether a stock is handled by NYSE American DMMs on the speed bump impact. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility.

Share is the percentage of trading volume of NSYE American versus the market. NBBO relevant is the percentage of quote records of NYSE American that are inside or at NBBO. NBBO setting is the percentage of quote records of NYSE American that are better than previous NBBO and defines the new NBBO. Cancel-to-trade is the ratio of canceled orders versus trades. These trading activity indicators are all measured in % terms. Quoted spread is the displayed bid-ask spread, Roll spread is implicit spread from trade price, while Effective spread is calculated from execution price against 60-second-before NBBO. All are liquidity indicators measured in percentage points. Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.

The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from July 10, 2017 to Aug 10, 2017. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

For stocks handled by specialist, NYSE Americans market share increases. It is more likely for those stocks to have quotes keeping up with NBBO on NYSE American. Roll spread and effective spread drops more for trades handled by specialists. Autocorrelation drops while information ratio increases. None of the distinction between specialist and non-specialist handled trades is statistically significant outside NYSE American. This implies the specialist on NYSE American benefit from the speed bump, which might be one of the motivations for NYSE to pursue this new trading system.

7 The impact of NYSE migration to Pillar

We've now seen two speed bumps that are both latency increasing trading regimes that slow the market down. How does the market performance change when the latency is reduced? Research has been done on the market quality change during the transition from traditional broker-dealer order book keeping to electronic trading, which is a huge market-wise latency reduction. For example, Hendershott, Jones & Menkveld (2011) show that the transition not only improves liquidity by significantly reducing the bid-ask spread. Hasbrouck (1995) discovers that electronic trading encourages more prompt price discovery as the general predictability of the equity market falls.

Since 2017, NYSE has been migrating its exchanges to a new platform, Pillar, with lower latency. In this section, We study NYSE Arca and NYSE National transition to Pillar, but find opposite results to previous research. There is no evidence that further latency reduction improves market quality.

NYSE Arca is an exchange specialized in trading Exchange Traded Products (ETP), such as ETF, ETN and ETV. With the rising market share of passive investing and index tracking, the market share of NYSE Arca is quite large. NYSE National, on the other hand, is relatively small. It is a pure trading platform that does not offer listing services, just like IEX (IEX successfully listed Interactive Brokers in Oct 2018, but the company went back to Nasdaq one year later.[37]) NYSE National also has similar market share compared to IEX.

[37] https://seekingalpha.com/news/3501066-iex-exit-listings-interactive-brokers-returns-nasdaq

7.1 Methodology

We use a diff-in-diff regression to investigate the impact of latency reduction policies:

$$y_{i,t} = \alpha_t + \delta MG_{i,t} + \beta MG_{i,t} * MarketShare_{i,t} + \gamma X_{i,t} + \eta MarketShare_{i,t} + \epsilon_{i,t} \tag{32}$$

where $y_{i,t}$ are market activity and market quality measures. α_t is the time (trading day) fixed effect. $MG_{i,t}$ is a binary variable indicating whether a stock has migrated on a particular day (and for all the days in the future). $MarketShare_{i,t}$ is the average trading volume NYSE Arca/ NYSE National market share quantile of a stock on a particular day. $MG_{i,t} * MarketShare_{i,t}$ is the interaction term that describes the diff-in-diff effect of latency reduction impact on high NYSE Arca/NYSE National market share stocks. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility. Controls will be calculated using 10-trading day average before the migration to Pillar for each stock.

The impact of the migration on two exchanges themselves and the entire market are both investigated. Similar to the case of NYSE American speed bump, migrations of NYSE National and NYSE Arca to Pillar were set up long ago. Thus, it is reasonable to consider these migrations as exogenous shocks rather than adaptive acts by NYSE.

7.2 Summary Statistics

The trading activities on NYSE Arca and NYSE National do not change much after their migrations to Pillar. We find that the price discovery and liquidity of NYSE Arca slightly improves. The price discovery on NYSE National does not change, while the liquidity is slightly worse after the migration. Both exchanges have increasing cancel-to-trade ratio, which is a proxy for the HFT activity. This is not surprising since a microsecond level latency reduction might not affect slower investors, but could be very important to a HFT.

Table 16: Equities traded on NYSE Arca or NYSE National: before and after migration to Pillar

Summary statistics	NYSE Arca		NYSE National	
	Before	After	Before	After
Market Share (%)	8.25	8.35	1.90	1.89
NBBO relevant (%)	13.6	13.1	11.7	12.1
Quoted Spread (bps)	55	49	54	49
Effective spread (bps)	37	40	38	41
Autocorrelation (60sec)	17.1	16.1	23.9	22.8
Autocorrelation (60sec)	16.4	16.7	15.1	14.9
Cancel-to-Trade (%)	27.6	37.1	35.5	40.1

This table covers all equities traded on NYSE Arca or NYSE National. The sample period covers 2 weeks (10 trading days) before the migration and runs from Aug 7, 2017 to Aug 20, 2017.

7.3 Results

NYSE Arca's migration increased liquidity a bit at the cost of price discovery. HFT activities seem to pick up, while the market share also increases a little.

Table 17: Diff-in-diff: The impact of NYSE Arca's migration to Pillar

Trading activity	NYSE Arca				Market	
	sniping	NBBO relevant	NBBO setting	Cancel-to-Trade	sniping	Cancel-to-Trade
Migration	1.31	3.43*	3.14	-3.53	-1.76	-3.65
	(1.4)	(1.6)	(0.6)	(-0.9)	(-1.5)	(-1.2)
Migration ×	-1.45*	0.54	2.43***	-1.54	-6.54	-2.42
Market share	(1.8)	(-1.4)	(7.4)	(-1.4)	(-0.8)	(-1.5)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.54	0.14	0.43	0.25	0.16	0.31

Liquidity	NYSE Arca			Market		
	price impact	Roll spread	Effective spread	price impact	Roll spread	Effective spread
Migration	-1.23	0.14*	0.32*	-2.15*	-3.22*	-1.04
	(-0.8)	(1.9)	(1.8)	(-1.6)	(-1.9)	(-1.1)
Migration ×	1.3*	0.54	0.34*	-1.35	1.23	-0.3
Market share	(1.6)	(1.1)	(1.8)	(-1.1)	(1.2)	(-1.3)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.24	0.25	0.16	0.17	0.42	0.27

Efficiency	NYSE Arca			Market		
	Autocorr(1sec)	Autocorr(60sec)	Information ratio	Autocorr(1sec)	Autocorr(60sec)	Information ratio
Migration	2.34	0.31*	0.12*	-4.5	2.1	-3.13**
	(1.5)	(1.8)	(1.9)	(-0.4)	(0.6)	(-2.5)
Migration ×	-1.10	0.23*	0.05	-0.31*	0.1	-1.73
Market share	(-0.7)	(1.8)	(0.6)	(-1.9)	(0.3)	(-1.5)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.17	0.58	0.15	0.31	0.37	0.44

This table shows market activity and quality change when NYSE Arca migrated to Pillar. We regress all measures of market activity and market quality on the speed bump implementation, market share of NYSE Arca of stocks, and characteristics of stocks:

$$y_{i,t} = \alpha_t + \beta MG_{i,t} + \delta MG_{i,t} * MarketShare_{i,t} + \gamma X_{i,t} + \eta MarketShare_{i,t} + \epsilon_{i,t} \quad (33)$$

where $y_{i,t}$ are variables in the first row of the table. α_t is the time (trading day) fixed effect. $MG_{i,t}$ is a binary variable indicating whether a stock has migrated on a particular day (and for all the days in the future). $MarketShare_{i,t}$ is the average trading volume market share quantile of a stock on a particular day. $MG_{i,t} * MarketShare_{i,t}$ is the interaction term that describes the diff-in-diff effect of NYSE Arca market share on the latency reduction impact. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.

Share is the percentage of trading volume of NYSE Arca versus the market. NBBO relevant is the percentage of quote records of NYSE Arca that are inside or at NBBO. NBBO setting is the percentage of quote records of NYSE Arca that are better than previous NBBO and defines the new NBBO. Cancel-to-trade is the ratio of canceled orders versus trades. These market activity indicators are all measured in % terms. Quoted spread is the displayed bid-ask spread, Roll spread is implicit spread from trade price, while Effective spread is calculated from execution price against 60-second-before NBBO. All are liquidity indicators measured in percentage points. Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.

The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from Aug 7, 2017 to Aug 31, 2017. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

NYSE National's migration does not change liquidity, and the price discovery drops a little. HFT activities seem to pick up as in the previous case, and the market share remained stable as well.

Table 18: Diff-in-diff: The impact of NYSE National's migration to Pillar

Trading activity	NYSE National				Market	
	sniping	NBBO relevant	NBBO setting	Cancel-to-Trade	sniping	Cancel-to-Trade
Migration	0.21	0.05*	0.25	-4.4*	-5.4	-1.43
	(1.4)	(1.6)	(0.8)	(-1.9)	(-0.7)	(-1.4)
Migration ×	0.16*	0.003	-0.01	-1.54	-1.65*	-1.37
Market share	(1.6)	(1.3)	(-0.4)	(-0.9)	(-1.7)	(-0.4)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.39	0.46	0.43	0.25	0.19	0.14

Liquidity	NYSE National			Market		
	price impact	Roll spread	Effective spread	price impact	Roll spread	Effective spread
Migration	-1.23***	0.17*	0.12*	-26.1	-21.2*	-1.04
	(-3.8)	(1.9)	(1.7)	(-1.2)	(-1.9)	(-0.6)
Migration ×	0.45*	0.23	-0.73*	-1.73	1.47	-0.3
Market share	(1.7)	(1.1)	(-1.8)	(-1.5)	(1.2)	(-1.3)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.23	0.36	0.56	0.43	0.44	0.13

Efficiency	NYSE National			Market		
	Autocorr(1sec)	Autocorr(60sec)	Information ratio	Autocorr(1sec)	Autocorr(60sec)	Information ratio
Migration	-0.12*	0.08	0.13	-0.31	0.12	-1.98*
	(-1.7)	(1.3)	(1.4)	(-0.2)	(0.7)	(-1.8)
Migration ×	-0.02	1.14*	2.52	-0.43	0.31	-4.73
Market share	(-1.5)	(1.7)	(1.5)	(-0.3)	(0.5)	(-1.0)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.53	0.33	0.24	0.34	0.19	0.22

This table shows market activity and quality change when NYSE National migrated to Pillar. We regress all measures of market activity and market quality on the speed bump implementation, market share of NYSE National of stocks, and characteristics of stocks:

$$y_{i,t} = \alpha_t + \beta MG_{i,t} + \delta MG_{i,t} * MarketShare_{i,t} + \gamma X_{i,t} + \eta MarketShare_{i,t} + \epsilon_{i,t} \quad (34)$$

where $y_{i,t}$ are variables in the first row of the table. α_t is the time (trading day) fixed effect. $MG_{i,t}$ is a binary variable indicating whether a stock has migrated on a particular day (and for all the days in the future). $MarketShare_{i,t}$ is the average trading volume market share quantile of a stock on a particular day. $MG_{i,t} * MarketShare_{i,t}$ is the interaction term that describes the diff-in-diff effect of NYSE National market share on the latency reduction impact. $X_{i,t}$ are a group of controls for a stock, for which we use market cap, volume and volatility, as shown in the table.

Share is the percentage of trading volume of NYSE National versus the market. NBBO relevant is the percentage of quote records of NYSE National that are inside or at NBBO. NBBO setting is the percentage of quote records of NYSE National that are better than previous NBBO and defines the new NBBO. Cancel-to-trade is the ratio of canceled orders versus trades. These market activity indicators are all measured in % terms. Quoted spread is the displayed bid-ask spread, Roll spread is implicit spread from trade price, while Effective spread is calculated from execution price against 60-second-before NBBO. All are liquidity indicators measured in percentage points. Autocorr(1sec) is average autocorrelation of stock price every second, while Autocorr(60sec) is average autocorrelation of stock price every minute. Information ratio is the percentage of information content versus trading noise. All are price efficiency indicators ranged from 0 to 1.

The sample period covers 2 weeks (10 trading days) on both sides of the event and runs from May 7, 2018 to May 31, 2018. */**/*** indicate statistical significance at the 90/95/99 percent levels, respectively. Standard errors are clustered by day.

NYSE National and NYSE Arca are two distinct exchanges. The former is small and focused on stock trading, while the latter has a large market share thanks to its edge on ETP trading. Neither of them shows evidence of market quality improvement with the reduction of latency. On the contrary, NYSE's further effort of reducing latency when migrating to the new trading platform Pillar seems to benefit no one but HFT according to our results.

One explanation is that latency is only one aspect of trading. Latency reduction comes with the transition to a more efficient and accessible electronic trading regime. However, latency reduction alone is not sufficient to further improve market quality, especially when the market is already very fast. Speed is costly to chase. Although it might be good for some participants, it is not necessarily good for the market, as we've seen in the cases of NYSE Arca and NYSE National.

8 Conclusion

Using daily TAQ data, this paper finds that including the IEX speed bump into the National Market System leads to better liquidity and better price discovery; that the speed bump discourages quote sniping and high-frequency order cancellation; and that the market quality improvements are larger for stocks heavily-traded in IEX, stocks with high volatility, and transactions happening closer to the IEX or the SIP. However, this paper finds that a seemingly identical speed bump by NYSE American does not induce market quality improvements, while only benefiting NYSE American designated market makers. This paper looks into an trading platform technology upgrade called "NYSE Pillar" for NYSE Arca and NYSE National, and finds that the overall latency reduction encourages HFT and strengthens data monopoly, but not improving market quality.

References

Aldrich, E. M. & Friedman, D. (2017), 'A theoretical model of the investors exchange'.

Aldrich, E. M. & Vargas, K. L. (2019), 'Experiments in high-frequency trading: comparing two market institutions', *Experimental Economics* pp. 1–31.

Angel, J. J., Harris, L. E. & Spatt, C. S. (2015), 'Equity trading in the 21st century: An update', *Quarterly Journal of Finance* **5**(01).

Aoyagi, J. (2020), 'The dark side of regulating fast informed trading', *Working paper* .

Aquilina, M., Budish, E. & O'Neill, P. (2022), 'Quantifying the high-frequency trading "arms race"', *The Quarterly Journal of Economics* **137**(1), 493–564.

Brown, A. & Yang, F. (2016), 'Slowing down fast traders: Evidence from the Betfair speed bump'.

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

Chen, H., Foley, S., Goldstein, M. A. & Ruf, T. (2017), 'The value of a millisecond: Harnessing information in fast, fragmented markets', *Fragmented Markets (November 18)* .

Chordia, T., Roll, R. & Subrahmanyam, A. (2000), 'Commonality in liquidity', *Journal of Financial Economics* **56**(1), 3–28.

Chung, K. H. & Chuwonganant, C. (2014), 'Uncertainty, market structure, and liquidity', *Journal of Financial Economics* **113**(3), 476–499.

Cumming, D., Johan, S. & Li, D. (2011), 'Exchange trading rules and stock market liquidity', *Journal of Financial Economics* **99**(3), 651–671.

Ernst, T., Sokobin, J. & Spatt, C. (2022), 'The value of off-exchange data'.

Gonçalves, J., Kräussl, R. & Levin, V. (2019), 'Do speed bumps prevent accidents in financial markets?', *Working paper* .

Hasbrouck, J. (1995), 'One security, many markets: Determining the contributions to price discovery', *Journal of Finance* **50**(4), 1175–1199.

Hendershott, T., Jones, C. M. & Menkveld, A. J. (2011), 'Does algorithmic trading improve liquidity?', *Journal of Finance* **66**(1), 1–33.

Hu, E. (2019), 'Intentional access delays, market quality, and price discovery: Evidence from IEX becoming an exchange'.

Jones, C. M. (2016), 'RE: File No.10-222, Investors Exchange, LLC (IEX) exchange application', *https://www.sec.gov/comments/10-222/10222-433.pdf* .

Khapko, M. & Zoican, M. (2019), 'Do speed bumps curb speed investment? evidence from a laboratory market', *Working paper* .

Lee, C. M. & Ready, M. J. (1991), 'Inferring trade direction from intraday data', *The Journal of Finance* **46**(2), 733–746.

Lee, T. (2019), 'Latency in fragmented markets', *Review of Economic Dynamics* .

Li, S., Ye, M. & Zheng, M. (2021), 'Financial regulation, clientele segmentation, and stock exchange order types'.

Shkilko, A. & Sokolov, K. (2020), 'Every cloud has a silver lining: Fast trading, microwave connectivity and trading costs', *Journal of Finance, Forthcoming* .

Spatt, C. S. (2019), 'Is equity market exchange structure anti-competitive?', *Working paper* .

Wah, E. & Feldman, S. (2018), 'Gone in sixty seconds: The cost of trading in long queues', *IEX Insights* .

Wang, X. (2018), 'Why do stock exchanges compete on speed, and how?'.

Zhu, J. (2019), 'A two-exchange model for HFT and the speed bump', *Working Paper* .

Chapter 2

Speed bump and high-frequency trading

1 Introduction

There have always been controversies about high-frequency trading (HFT). Some investors, such as The Vanguard Group[1] and AQR Capital Managment,[2] welcome HFT because they believe it helps lower transaction costs. Others (e.g. Charles Schwab)[3] claim that high-frequency traders have an unfair advantage over slow traders and should be regulated. Some research shows HFT is good for price discovery (Brogaard et al., 2014). Others argue that too much HFT activity could prevent new information from being produced for the market (Weller, 2016).

The advantage of being able to trade faster comes from two facts. First, almost all public exchanges use continuous double auctions, or CDA, as the trading protocol. Under this protocol, even an infinitely small fraction of a second provides the trader time priority. Second, slow traders are more vulnerable in fragmented markets where quote-matchers can front-run orders in one market by trading in another market (Angel, Harris & Spatt, 2015). Researchers have proposed several alternative trading protocols to address this issue, a "frequent batch auction"(Budish, Cramton and Shim, 2015), where orders are processed discretely (e.g. every tenth of a second). However, so far there have been no implementation of a batch auction trading protocol.

[1] https://www.cbsnews.com/news/jack-bogle-michael-lewis-is-wrong-about-rigged-markets/
[2] https://www.wsj.com/articles/high-frequency-hyperbole-1396394601
[3] https://blogs.wsj.com/moneybeat/2014/04/03/schwab-on-hft-growing-cancer-that-must-be-addressed/

An alternative approach is to setup an artificial "speed bump" to trading. Specifically, all inbound and outbound trading messages are delayed by 350 microseconds by forcing them through a 38-mile cable that is coiled in a box. It was first proposed by The Investors' Exchange LLC (IEX) in 2013. The concept of "speed bump" has become much more well-known since June 17th, 2016, when United States Securities and Exchange Commission (SEC) approved IEX operating as a public stock exchange.[4]

As was depicted in Michael Lewis's "Flash Boys", IEX was established with the goal of giving investors protection against high-frequency traders. In the press release, SEC said it approved IEX's application to launch a national exchange because the Commission believes this action would "promote competition and innovation," so that the market could "continue to deliver robust, efficient service to both retail and institutional investors."

However, there's been much debate on whether a speed bump will help IEX achieve these goals. For example, Prof. Charles M. Jones wrote a public letter to SEC (March 2, 2016), expressing his concern of IEX running as a public exchange. He argued that a speed bump, if implemented, would be far from de minimis in trading and could potentially harm market liquidity. Similar concerns have been raised when the Canadian exchange TSX Alpha implemented a speed bump but allowed investors to pay for a specific type of order that bypass the artificial delay. Chen et al.(2017) find that the asymmetric speed bump segments order flow and increases profits for fast liquidity providers at the expense of other liquidity providers and aggregate market quality.

Since IEX's speed bump debuted, other exchanges have followed suit. NYSE American has a very similar speed bump implemented since July 2017.[5] Cboe is seeking to introduce a brief delay on EDGA.[6] Investors are now in a market where they can choose to trade on an exchange with or without a speed bump or without.

In this paper, we study the actions and interactions of different market participants when they have that choice. We find that a speed bump provides protection to the fundamental investor against front-

[4]https://www.sec.gov/news/pressrelease/2016-123.html
[5]https://www.nyse.com/pillar
[6]https://www.wsj.com/articles/new-speed-bump-planned-for-u-s-stock-market-1535713321

running, increasing his profitability at the expense of the HF trader. When the fundamental investor trade more aggressively, price discovery is improved while liquidity decreases. Each market participant's trading strategy depends on information production, communication between market makers, as well as venue choice of uninformed investors. The impact of introducing a speed bump varies under different market conditions.

1.1 Related literature

This paper is related to five branches of literature: The Kyle model and its extensions; latency models; impact of high-frequency trading on information acquisition and speed technology purchase; impact of high-frequency trading on market liquidity and price informativeness; and research about speed bump regulation.

Kyle (1985) is germane to theoretical models of trading on private information for profit. In an extended Kyle framework, Li (2014) finds that front-running by high-frequency traders effectively levies a speed tax on normal-speed traders, and its negative effects on market quality are more severe when high frequency traders have more heterogeneous speeds. Yang and Zhou (2016) add "order-flow informed" investors to a two-period Kyle model, investigating the strategic interaction of seeking and hiding fundamental information. Kumar and Seppi (1994) study the role of index arbitrage in the transmission of information across markets and determinants of pricing rules and optimal trading strategies in a Kyle setting. In this paper, we model investors' behavior when there are multiple rounds of trading and two markets with different messaging latencies.

Two latency papers are most relevant for this paper. Wang (2018) explores how and why exchanges compete on order processing speeds as a service appealing to fast traders who take advantage of the Order Protection Rule. Lee (2018) uses the Kyle model to explore how a symmetric cross-venue latency affects informed traders' optimal strategy, where a larger latency makes cross-venue information less credible. Our model has both informed traders and high-frequency arbitrageurs. A speed bump has very different impact on them.

Studies on the impact of high-frequency trading can be roughly divided into two groups. The first group

views this issue from a social planner's perspective, and cares about efficient resources allocation. Weller (2016) finds evidence that algorithmic trading strongly deters information acquisition despite its power in translating available information into prices. Budish, Cramton and Shim (2015) argue that too much has been invested to reduce trading latency, and propose an alternative trading scheme to discourage this technological "arms race." The model in this paper aims to characterize both information production and price informativeness by investigating the impact of a speed bump on the trading behavior of fundamental investors and high-frequency traders.

The second group asks questions on behalf of participant traders, such as: do high-frequency traders provide additional liquidity? Does the price contain more information when there are high-frequency traders around? Madrigal (1996) shows that "non-fundamental" speculators can lead to reductions in liquidity and the information content of prices, even in an efficient market. Farboodi and Veldkamp (2017) set up a long-run growth model to understand how improvements in financial technology shape information choices, trading strategies and market efficiency. Tong (2015) finds empirical evidence that although high-frequency trading decreases the average bid-ask spread, it increases traditional institutional traders' overall trading cost by imposing execution shortfall. In our paper, we will look at the impact of a speed bump on market liquidity.

There are a few studies on speed bump regulation. Using data from a betting exchange, Brown and Yang (2016) find evidence that the speed bump in Betfair protects slower traders and that fast traders develop strategies to circumvent the speed bump. Aldrich and Friedman (2017) consider three order types in particular: primary peg, midpoint peg and discretionary peg, which are unique to IEX. The detailed model of the price formation mechanism in a continuous limit order book market generates a set of predictions that allows for comparison with traders' behavior in a traditional market, as well as testing in the field. Using TAQ data, Hu (2018) finds that the introduction of a speed bump in IEX helps the overall price discovery and liquidity on the market. Our paper has a broader definition for a speed bump, so the implications would apply not only to IEX, but also to other exchanges with regulatory innovations with the purpose of deterring predatory high-frequency front-running.

The rest of this paper proceeds as follows. The model in Section II compares the behavior of different market participants before and after we introduce a speed bump. Section III characterizes solutions to the model and discusses some of its implications. We extend the model in Section IV to consider alternative assumptions. The extensions serve not only as robustness checks of the results we have in Section III, but tells us how the impact of a speed bump varies under different market conditions. Section V concludes.

2 Baseline model

In this section, we set up a baseline model to explain the trading mechanism, price formation, and the interaction between fundamental investors and high-frequency traders in fragmented markets.

This is a one-period model with multiple rounds of trading. There is only one single asset for trading. The liquidation value of the asset v has a distribution $N(v_0, \sigma_v^2)$. The asset is traded on two exchanges, A and B, which are identical except that their servers (order matching engines) are in different locations. Market structure and distribution of the asset value are common knowledge to all market participants.

2.1 Trading time-line

There are three rounds of trading. In the beginning, a fundamental investor receives a signal s of the value of the asset v. Based on that signal, he decides the volume he submits to both exchanges, x^A and x^B. Depending on the location of the broker (or Securities Information Processor, SIP) he uses, orders will arrive at one exchange earlier. Without loss of generality, assume he routes his order so that it arrives at exchange A first. The uninformed traders' orders arrive at the same time, with signed amount u_1 that has distribution $N(0, \sigma_u^2)$. The market maker at exchange A sets the price p_1^A based on aggregate order flow $x^A + u_1$.

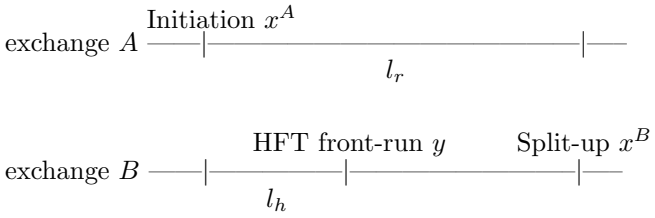
The high-frequency trader observes this first round of trading and extracts information from aggregate order flow $x^A + u_1$. He then decides the signed amount y he submits to exchange B. Due to an inter-

market lag, the market maker at exchange B does not know of the trade on exchange A yet. Similarly, the uninformed traders' orders arrive at the same time, with signed amount u_2 that has distribution $N(0, \sigma_u^2)$. The market maker at exchange B observes aggregate order flow $y + u_2$ and sets price p_2^B.

The third round of trading happens when order x^B arrives at exchange B. Again, the uninformed traders' orders arrive at the same time, with signed amount u_3 that has distribution $N(0, \sigma_u^2)$. Considering both p_2^B and $x^B + u_3$, the market maker at exchange B sets the price p_3^B.

In the end, the liquidation value is realized by both the HF trader and the fundamental investor.

The trading time-line when there is no speed bump is summarized in the following chart:



2.2 Latencies and the speed bump

To better illustrate the impact of the speed bump, this section introduces two key latencies of electronic trading:

- Routing latency l_r: This is the time it takes a fundamental investor's order to travel from exchange A to B through the public SIP. When the fundamental investor submits his order to both exchanges, this latency is the time difference between when those orders arrive at closer exchange and more distant one.

- HF trader communication time l_h: This is the time it takes the HF trader to process the order-flow information in one exchange, decide how to react, and send an order to another exchange. Since the HF trader buys co-location service from exchanges and builds his own high-speed network, $l_h < l_r$,

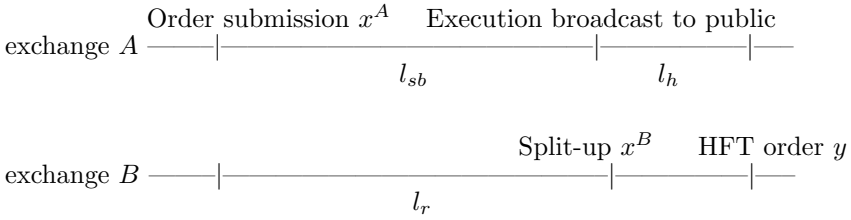
that gives the HF trader the ability to front-run.

In this paper, a speed bump is modeled as an artificial transmission delay, l_{sb}, for orders in and out of a certain exchange, say A. It takes longer to process orders submitted to exchange A. When a trade happens, it takes longer for traders outside exchange A to know of that trade.

When the fundamental investor chooses to be closer to exchange B, the speed bump on exchange A won't change the sequence of trading. However, when he is closer to exchange A, a speed bump would prevent the HF trader from front-running the fundamental investor's orders as long as $l_{sb} + l_h > l_r$. In this case, trading time-line would be:

1. The fundamental investor submits orders to both exchange A and B
2. Order x^A arrives at exchange A and is executed. Because of the speed bump, the HF trader learns that trade with a delay l_{sb}.
3. Rest of the fundamental investor's order arrives at exchange B
4. The HF trader's order arrive at exchange B

The trading time-line when there is a speed bump on exchange A is summarized in the following chart:



Notice that now the fundamental investor's orders arrive at exchange B faster than the HF trader's, thanks to the speed bump in exchange A. This gives the fundamental investor some advantage, as we will see in later sections.

2.3 Equilibrium

An equilibrium consists of the trading strategy $X = (x^A, x^B)$ of the fundamental investor, the trading strategy y of the HF trader, and the pricing strategy $P = (p_1^A, p_2^B, p_3^B)$ of the two local market makers.

When there is no speed bump:

The fundamental investor maximizes profit:

$$\max_{x^A, x^B} r_f = x^A(v - p_1^A) + x^B(v - p_3^B)$$

The HF trader maximizes profit:

$$\max_{y} r_h = y(v - p_2^B)$$

The market makers set competitive price incorporating all available information

$$p_1^A = E[v|x^A + u_1]$$

$$p_2^B = E[v|y + u_2]$$

$$p_3^B = E[v|x^B + u_3, p_2^B]$$

When there is a speed bump on exchange A:

The fundamental investor maximizes profit:

$$\max_{x^A, x^B} r_f = x^A(v - p_1^A) + x^B(v - p_2^B)$$

The HF trader maximizes profit:

$$\max_{y} r_h = y(v - p_3^B)$$

The market makers set competitive prices incorporating all available information

$$p_1^A = E[v|x^A + u_1]$$

$$p_2^B = E[v|x^B + u_2]$$

$$p_3^B = E[v|y + u_3, p_2^B]$$

In both cases, there exists an equilibrium $X = (x^A, x^B), y, P = (p_1^A, p_2^B, p_3^B)$. Define the signed amounts of aggregate orders for the three rounds of trading as z_1, z_2, z_3. Specifically, the functional forms for equilibrium prices and trading strategies are all linear:

$$x^A(s) = \alpha\beta(s - v_0)$$

$$x^B(s) = (1 - \alpha)\beta(s - v_0)$$

$$y(z_1) = \gamma(z_1)$$

$$p_1^A = v_0 + \lambda_1(z_1)$$

$$p_2^B = v_0 + \lambda_2(z_2)$$

$$p_3^B = p_2^B + \lambda_3(z_3)$$

The equilibrium $\{X, y, P\}$ is determined by the exogenous parameters $v_0, \sigma_v, s, \sigma_u$.

3 Discussion

Now that we have the equilibrium, we first characterize all market participants' behavior in the model, and then discuss some of the important measures of market quality. In the following discussions, the superscript ^{S} stands for the solution with the speed bump.

3.1 The fundamental investor's decision

A fundamental investor's decision consists of two parts: how aggressive (β) he trades on a piece of information (signal s); how he splits the order between two exchanges (α).

A speed bump protects the fundamental investor from being exploited by the HF trader, since both rounds of their orders will be executed before the HF trader response. As a result, they are more comfortable trading higher volumes ($\beta^S > \beta$) and revealing more information to the market.

Meanwhile, thanks to the speed bump, the fundamental investor worries less about first round of trading revealing information that negatively affects second round of trading, since the HF trader no longer front-runs. In other words, there is no need to concentrate orders earlier. Thus, lower portion ($\alpha^S < \alpha$) of the total order is submitted to the closer exchange A.

Note that here a speed bump works opposite ways on these two parameters: $\beta^S > \beta, \alpha^S < \alpha$. Thus, whether a fundamental trader trades more in absolute amount at the closer exchange A is not fully determined:

$$|\alpha^S \beta^S (s - v_0)| > |\alpha\beta(s - v_0)| \iff \alpha^S \beta^S > \alpha\beta$$

However, we know for sure he would trade more at exchange B:

$$|(1 - \alpha^S)\beta^S (s - v_0)| > |(1 - \alpha)\beta(s - v_0)|$$

3.2 The HF trader's arbitrage

The HF trader only decides how aggressive they should trade (γ) on the order-flow information from the first round of trading. They consider two things: whether they trade before or after the fundamental trader's order; and how much they trust the volume information. The former depends on whether there is a speed bump. The latter is closely related to how the order-flow information transfers from one exchange to another.

The HF trader's prediction of true value of the asset improves since they observe the first round of trading. We measure it using the variance reduction after that piece of information is revealed to the HF trader:

$$\Delta I_H = Var[v] - Var[v|p_1^A]$$

Thus, for unit volume of trading by the informed investor in the first round, the HF trader gets information:[7]

$$\gamma = \frac{\sigma_v^2 - Var[v|p_1^A]}{E[|z_1|]}$$

When the above ratio is higher, the HF trader dares to bet more on a certain volume they observe (bigger γ), because the more information contained in unit volume, the more accurate the signal extracted from order-flow, and the higher the profitability of arbitrage. Solving two equilibriums we know $\gamma^S < \gamma$.

3.3 Market maker's pricing strategy

There are three rounds of pricing $(\lambda_1, \lambda_2, \lambda_3)$ by two local market makers. The market maker in exchange A only cares about how informative the first round of trading is:

$$\Delta I_A = \sigma_v - Var[v|z_1]$$

Now if $\alpha^S \beta^S > \alpha\beta$, then $E[|z_1^S|] > E[|z_1|]$, volume of the first round is more informative, $\lambda_1^S > \lambda_1$, and vice versa.

The market maker in exchange B sets the prices for the second and third round of trading. When there is no speed bump, the second round is from the HF trader:

$$\Delta I_{B,H} = \sigma_v - Var[v|\gamma(\alpha\beta(s - v_0) + u_1) + u_2]$$

[7]For a Gaussian random variable $X \sim N(\mu, \sigma^2)$, the mean of its absolute value, $Y = |X|$, is $\mu_Y = \sigma\sqrt{\frac{2}{\pi}}e^{-\frac{\mu^2}{2\sigma^2}} + \mu(1 - 2\Phi(-\frac{\mu}{\sigma}))$

and third round is from the fundamental trader:

$$\Delta I_{B,F} = Var[v|\gamma(\alpha\beta(s - v_0) + u_1) + u_2] - Var[v|(1 - \alpha)\beta(s - v_0) + u_3]$$

However, when there is a speed bump, the second round is from the fundamental trader:

$$\Delta I^S_{B,F} = \sigma_v - Var[v|(1 - \alpha^S)\beta^S(s - v_0) + u_2]$$

and third round is from the HF trader:

$$\Delta I^S_{B,H} = Var[v|(1 - \alpha^S)\beta^S(s - v_0) + u_2] - Var[v|\gamma^S(\alpha^S\beta^S(s - v_0) + u_1) + u_3]$$

Comparing them we get $\Delta I^S_{B,F} > \Delta I_{B,F}, \Delta I^S_{B,H} < \Delta I_{B,H}$. That is, in exchange B, the market maker sets the price more aggressively for the fundamental investor with the speed bump, and less aggressively for the HF trader.

3.4 Price discovery

For each round of trading, price discovery is defined as $Var[v|p_k], k = 1, 2, 3$. To see the overall price discovery of the market, we calculate the expectation of the asset's liquidation value conditional on all information from both exchanges:

$$p_d = E[v|p^B_3, p^A_1]$$

Table 1: Price discovery

Variable	Change after speed bump:$\alpha^S\beta^S > \alpha\beta$	Change after speed bump:$\alpha^S\beta^S < \alpha\beta$
$Var[v\|p^A_1]$	decrease	increase
$Var[v\|p^B_2]$	decrease	decrease
$Var[v\|p^B_3]$	increase	decrease
$Var[v\|p_d]$	decrease	decrease

Overall price discovery improves because the fundamental investor is protected by the speed bump and trade more aggressively, bringing more information to the market. Price discovery in exchange A increases only when the fundamental investor trades more in the first round. The first round price discovery in exchange B increases because better-informed fundamental investor is now ahead of the HF trader when there is a speed bump. Second round price discovery in exchange B is the complement of price discovery in exchange A, and will certainly go up if price discovery in exchange A decreases. Price discovery improvements can happen in either exchange depending on parameterization of the model, but overall price discovery always improves.

3.5 Liquidity

We follow Kyle's definition to measure liquidity as $q_k = \frac{1}{\lambda_k}$, where $k = 1, 2, 3$, the price impact of trading. We also calculate the volume-weighted market wide liquidity:

$$q = \sum_{k=1}^{3} \frac{E[|z_k|]}{E[|z_1| + |z_2| + |z_3|]} \frac{1}{\lambda_k}$$

where $|z_k|, k = 1, 2, 3$ denotes the absolute value of the volume of each of the 3 rounds of trading.

Table 2: Liquidity

Variable	Change after speed bump:$\alpha^S\beta^S > \alpha\beta$	Change after speed bump:$\alpha^S\beta^S < \alpha\beta$
q_A	decrease	increase
$q_{B,F}$	decrease	decrease
$q_{B,H}$	increase	increase
q	decrease	decrease

When a market maker anticipates the protected fundamental investor will trade more aggressively, the price impact of the trade increases, and thus liquidity decreases. It's always true for trading at exchange B:

$$|(1 - \alpha^S)\beta^S(s - v_0)| > |(1 - \alpha)\beta(s - v_0)|$$

At exchange A, it is possible that when $\alpha^S\beta^S < \alpha\beta$, price impact is smaller and the liquidity improves. However, this only happens with a very small volume. Both this and the improved liquidity for the round where the HF trader trades with uninformed traders are dominated by the fact that liquidity is worse because of aggressive trading by fundamental investors.

3.6 Profitability

The HF trader only engages in one round of trading, thus his profit is easy to calculate. For the fundamental investor, we decompose his profit into two parts: initiation round r_i, and split-up round r_s, where $r_i + r_s = r_f$. Since the market maker is break-even in expectation, the total expected profits gained by the HF trader and the fundamental investor is the loss incurred to uninformed investors:

$$r_u = -(r_f + r_h)$$

Table 3: Profitability

Variable	Change after speed bump:$\alpha^S\beta^S > \alpha\beta$	Change after speed bump:$\alpha^S\beta^S < \alpha\beta$
r_i	increase	decrease
r_s	increase	increase
r_f	increase	increase
r_h	decrease	decrease
r_u	decrease	increase

There is a key trade-off for the fundamental investor's profit: more aggressive trading gains higher profit but reveals more information that negatively affects pricing of the follow-up round. Here with the help of speed bump, the gain of aggressive trading dominates. Although the HF trader now has better guessing of the true information, his profit drops since that piece of information is too late.

3.7 Cross-sectional variation: volatility

This section looks at the speed bump impact when the asset has a higher volatility, then compares it with the baseline model in the previous section. The model predicts that the protection of a speed bump is stronger for more volatile assets.

In the baseline model, asset v has a distribution $N(v_0, \sigma_v^2)$. For ease of notation, in this section we denote it as $v_L \sim N(v_0, \sigma_L^2)$ In addition, we assume asset v_H has the same mean v_0, but with a higher variance $\sigma_H > \sigma_L$, so $v_H \sim N(v_0, \sigma_H^2)$. We solve for the equilibrium condition for the high volatility case, and look at three key comparative statics:

- The price impact change after the speed bump: $\Delta\lambda_k|\sigma_H < \Delta\lambda_k|\sigma_L$

  Price impact after the speed bump increases less for v_H, implying that liquidity after the speed bump is better for v_H.

- The conditional variance change after the speed bump: $\Delta Var[v|p_d; \sigma_H] > \Delta Var[v|p_d; \sigma_L]$

  Conditional variance of the asset price after the speed bump drops more for v_H, indicating that more information is revealed and the price discovery is better for v_H.

- The informed trader's expected return change after the speed bump: $\Delta r_f|\sigma_H > \Delta r_f|\sigma_L$

  The informed trader's expected return after the speed bump increases more for v_H, meaning that the informed trader enjoys better protection from the speed bump when the asset has a higher volatility.

Thus, the model predicts that the speed bump has a larger impact for more volatile stocks.

4 Extended model

In this section the baseline model is extended in three different ways. First, we let information production be an endogenous decision made by the fundamental investor. Second, two local market makers are allowed to communicate with each other. Third, the uninformed traders have the discretion to choose their trading

venues and/or allocate orders between two exchanges.

We use the idea of difference-in-difference to explain the impact of a speed bump with different extensions. The differences of key variables before and after the speed bump will be listed out; we also show how large those changes compare against changes in baseline model. The latter tells us how the impact of speed bump varies under different market conditions. Specifically, for all key variable Ω, we calculate its equilibrium values under four environments:

1. $\Omega_{0,0}$: Baseline model, no speed bump
2. $\Omega_{0,1}$: Baseline model, with speed bump
3. $\Omega_{1,0}$: Extended model, no speed bump
4. $\Omega_{1,1}$: Baseline model, with speed bump

Then $\Omega_{0,1} - \Omega_{0,0}$ is the impact of speed bump in the baseline case, as we've already covered;

$\Omega_{1,1} - \Omega_{1,0}$ is the impact of a speed bump in the extended model, as will be shown in the second columns of tables in following sections.

$|(\Omega_{0,1} - \Omega_{0,0})/\Omega_{0,0}|$ is the magnitude of change in the baseline case, while $|(\Omega_{1,1} - \Omega_{1,0})/\Omega_{1,0}|$ is the magnitude of change in the extended model. We compare them and show the results in the third columns of tables in following sections.

4.1 Market maker communication

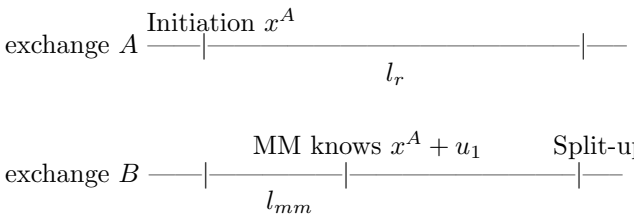
Until now we've assumed market makers only use local information to price the asset. What if the market maker observes what happens at both exchanges? To answer that question, we first need to define another latency:

- The market maker latency l_{mm}: This is the time it takes the global market maker to receive and digest order-flow information from the other exchange

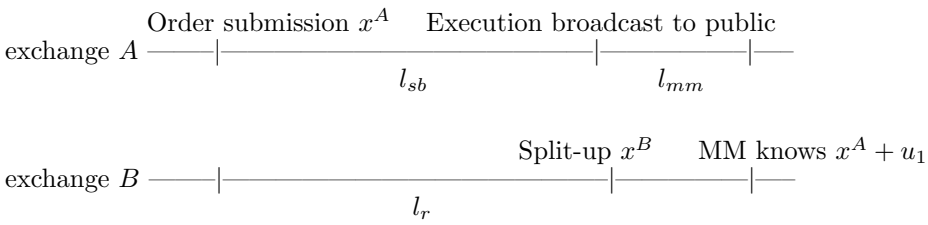
Depending on how fast the market maker can communicate, the trading landscape can be very different.

Case $l_{mm} \leq l_h$: There will be no arbitrage chance for the HF trader in this case, since the market maker has equal or even better speed technology as the HF trader. Implicitly we have $l_h < l_r$, thus pricing at the more distant exchange will base on the price of the first round trading, which affects the fundamental investor negatively as well.

When there is no speed bump, time-line of trading is as follows:



With a speed bump, the fundamental investor is protected from front-running not from the HF trader, but from the market maker.



This situation is summarized in the following table. The first column lists variables of interest. The second column shows the change due to speed bump in the extended model. The third column compares the change in the extended model and the change in the baseline model:

- '+' means they are in the same direction
- '–' means they are in the opposite direction
- 'bigger' means magnitude of change (as defined previously) in the extended model is bigger
- 'smaller' means magnitude of change in the extended model is smaller.

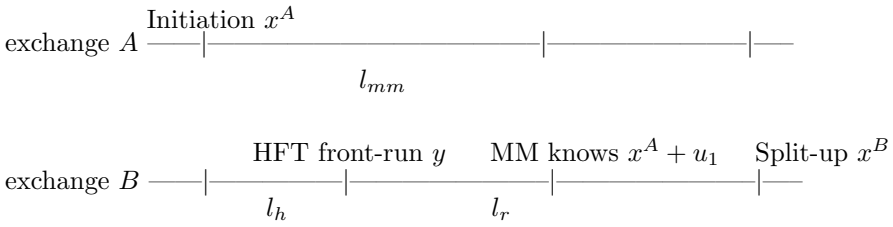
Table 4: Market Maker Communication: fast speed

Variable	Change after speed bump	Change compared to baseline
$Var[v\|p_1^A]$	increase	+,smaller
$Var[v\|p_2^B]$	increase	+,bigger
$Var[v\|p_3^B]$	decrease	+,bigger
$Var[v\|p_d]$	increase	+,bigger
q_1	increase	–,smaller
q_2	increase	+,bigger
q_3	increase	+,smaller
q	increase	–,smaller
r_i	increase	+,bigger
r_s	decrease	+,bigger
r_f	increase	+,smaller
r_h	decrease	+,smaller
r_u	increase	–,smaller

In this case, the HF trader is non-existent. Lee (2018) has a similar setup: as cross-venue latency decreases, liquidity and price discovery improve while the expected profits of informed traders decline. His findings coincide with our results here.

Case $l_r > l_{mm} > l_h$: The market maker is not fast enough to know the first round of trading before the HF trader front-runs. However, the market maker is faster than order routing of the fundamental investor, and thus pricing at the more distant exchange B will incorporate the information of the first round of trading in exchange A.

When there is no speed bump, time-line of trading is as follows:



When there is a speed bump, the extended model will be the same as the baseline model, since the speed bump prevents the market maker from learning what happens at exchange A in a timely manner.

This situation is summarized in the following table:

Table 5: Market Maker Communication: slow speed

Variable	Change after speed bump	Change c compared to baseline
$Var[v\|p_1^A]$	increase	+,smaller
$Var[v\|p_2^B]$	increase	+,smaller
$Var[v\|p_3^B]$	decrease	–,bigger
$Var[v\|p_d]$	increase	+,smaller
q_1	increase	–,smaller
q_2	increase	+,bigger
q_3	increase	+,bigger
q	increase	–,smaller
r_i	increase	+,smaller
r_s	decrease	+,bigger
r_f	increase	+,bigger
r_h	decrease	+,bigger
r_u	increase	–,smaller

Slow speed market maker communication improves the liquidity at the expense of a slightly worse price discovery. Profitability of the uninformed investors increases, while the fundamental investor and HF trader are not affected so much.

Case $l_{mm} > l_r$: The market maker is too slow to respond in time to what happens at the other exchange. This is the same as the baseline case.

4.2 Endogenous information production

Now the signal of liquidation value of the asset is not an endowment anymore. The fundamental investor has certain technology and can decide how much to invest for a more accurate signal. Specifically, when the fundamental investor pays the cost c, volatility of the signal becomes:

$$\sigma_s = e^{-c}\sigma_v$$

A more accurate signal allows the fundamental investor to trade more aggressively.

This situation is summarized in the following table.

Table 6: Endogenous information production

Variable	Change after speed bump	Change compared to baseline
$Var[v\|p_1^A]$	increase	–,smaller
$Var[v\|p_2^B]$	decrease	+,bigger
$Var[v\|p_3^B]$	decrease	–,smaller
$Var[v\|p_d]$	increase	+,smaller
q_1	decrease	+,bigger
q_2	increase	+,smaller
q_3	increase	–,bigger
q	increase	+,smaller
r_i	decrease	–,smaller
r_s	increase	+,bigger
r_f	increase	+,smaller
r_h	decrease	+,bigger
r_u	increase	–,smaller

The endogenization of information production intensifies the trade-off: the market enjoys even better price discovery, and has to bear with much worse liquidity. The fundamental investor's profitability improves even further at the expense of both the HF trader and the uninformed trader.

4.3 Uninformed investors with discretion

Three cases of discretion for the uninformed investors are discussed in this section. In the first case, they can choose the amount to trade, but within a certain exchange. In the second case, they are allowed to allocate a pre-specified amount between two exchanges. In the final case, both constraints are relaxed, where they can allocate whatever amount of orders to wherever they wish to trade.

1. Price-sensitive uninformed investors: Suppose the uninformed investors have the utility function:

$$U(e;\tau) = -\alpha(\tau - e)^2 - \lambda e^2$$

where τ is the pre-determined optimal quantity of trading. λ is a measure of price impact and is defined as in previous equilibriums. α measures how much uniformed investors hate not being able to trade their ideal amount relative to price impact loss. e is the actual amount of trading and can now be determined by

uniformed investors. Notice it could be the case that an uninformed investor is engaged in two rounds of trading at one exchange. We make the simplified assumption that the dis-utility from different rounds can be added:

$$U(e_1, e_2; \tau) = -\alpha(\tau - e_1 - e_2)^2 - \lambda_1 e_1^2 - \alpha(\tau - e_2)^2 - \lambda_2 e_2^2$$

Here we still assume that the uninformed investors at the two exchanges are separate.

Table 7: Uninformed traders' price sensitiveness

Variable	Change after speed bump	Change compared to baseline
$Var[v\|p_1^A]$	increase	+,bigger
$Var[v\|p_2^B]$	increase	+,smaller
$Var[v\|p_3^B]$	decrease	+,bigger
$Var[v\|p_d]$	increase	+,smaller
q_1	increase	+,bigger
q_2	decrease	+,smaller
q_3	decrease	+,bigger
q	increase	+,smaller
r_i	decrease	+,smaller
r_s	decrease	+,bigger
r_f	decrease	+,smaller
r_h	increase	+,smaller
r_u	increase	+,bigger

Letting the uninformed traders choose the volume of trading does not change the essence of the results from the baseline model. The uninformed investors relative disadvantageous position in trading is more because they are trenched than they are not free to choose trading amount.

2. Order allocation between two exchanges: The uniformed investors can use their discretion in an alternative way by allocating pre-determined quantity of trading between two exchanges. In the baseline case analysis, we see that uninformed investors are better off when they trade in exchange B (the exchange without a speed bump). Thus, it would be interesting to see what would the equilibrium look like when they are allowed to allocate their volume among different rounds of trading.

Uninformed traders' venue choice won't affect sequence of trading, but has any impact on the market makers' pricing strategy and trading strategies of other traders. They still get dis-utility from the price impact, but strategical allocate the trading amount so that total amount is equal to the pre-determined

quantity:

$$e^A + e_1^B + e_2^B = \tau$$

$$U(e^A, e_1^B, e_2^B; \tau) = -\lambda^A (e^A)^2 - \lambda_2^B (e_1^B)^2 - \lambda_2^B (e_2^B)^2$$

Here w.o.l.g. we assume there are one round of trading on exchange A and two rounds of trading on exchange B. This situation is summarized in the following table.

Table 8: Uninformed traders' venue choice

Variable	Change after speed bump	Change compared to baseline
$Var[v\|p_1^A]$	increase	+,bigger
$Var[v\|p_2^B]$	increase	–,smaller
$Var[v\|p_3^B]$	decrease	+,bigger
$Var[v\|p_d]$	increase	+,smaller
q_1	increase	–,bigger
q_2	decrease	+,smaller
q_3	decrease	+,bigger
q	increase	–,smaller
r_i	decrease	+,smaller
r_s	decrease	+,bigger
r_f	decrease	–,smaller
r_h	increase	–,smaller
r_u	increase	+,bigger

Venue discretion by the uninformed traders weakens the effect of price discovery improvements from the speed bump. Their venue choice improves both the market liquidity and their own profitability.

3. A comprehensive look at both volume and venue choice: The model in this section takes into consideration both of the above two variations of uniformed trading. Suppose the first group of uniformed investors are trenched in one particular exchange, but they are price-sensitive and can choose optimally the amount to trade. The second group of uniformed investors have fixed trading needs but can trade anywhere they like. Together they form the uninformed side of trading and the aggregate trading quantity would affect pricing in each exchange. Here w.o.l.g. we assume A has speed bump while B does not, and there are one trading in A and two tradings in B.

First, we define the utility of trenched, price sensitive uninformed innovators in exchange A:

$$U_{PS}^{A}(e^{A};\tau) = -\alpha(\tau - e^{A})^2 - \lambda^{A}(e^{A})^2$$

Second, we define the utility of trenched, price sensitive uninformed innovators in exchange B:

$$U_{PS}^{B}(e_1^{B}, e_2^{B};\tau) = -\alpha(\tau - e_1^{B})^2 - \lambda_1^{B}(e_1^{B})^2 - \alpha(\tau - e_2^{B})^2 - \lambda_2^{B}(e_2^{B})^2$$

Finally, we define the utility of venue choice uniformed investors that trade fixed aggregate amount between exchange A and B:

$$U_{VC}(i^{A}, i_1^{B}, i_2^{B};\tau) = -\lambda^{A}(i^{A})^2 - \lambda_2^{B}(i_1^{B})^2 - \lambda_2^{B}(i_2^{B})^2$$

where

$$i^{A} + i_1^{B} + i_2^{B} = \tau$$

For each round of trading, we know that

$$u^{A} = e^{A} + i^{A}$$

$$u_1^{B} = e_1^{B} + i_1^{B}$$

$$u_2^{B} = e_2^{B} + i_2^{B}$$

The rest of the equilibrium naturally follows from the baseline case we previous discussed.

Table 9: comprehensive look

Variable	Change after speed bump
u^A	decrease
u_1^B	increase
u_2^B	increase
e^A	decrease
e_1^B	decrease
e_2^B	increase
i^A	decrease
i_1^B	increase
i_2^B	increase
U_{PS}^A	decrease
U_{PS}^B	increase
U_{VC}	decrease

When both amount discretion and venue choice are considered, the uninformed investors are at a even better position. Market liquidity improves because now they are willing to trade in a certain round even if that specific round does not provide a very high utility, as long as they expect that trading would incur profitable trading later.

5 Conclusion

In this paper we build a model where all investors trade in two exchanges. When we introduce a speed bump to one of them, overall price discovery improves while market has lower liquidity. Fundamental investors become more profitable at the expense of HF traders, while uninformed investors can be better or worse. Using the comparative statics analysis, this paper finds that speed bump impact is stronger with higher volatility assets. This paper extends the model and finds that communication between market makers, endogenous information production, and uninformed investor's discretion all change liquidity and slow traders profit.

The results here have several empirical implications. The model predicts that both changing trading environment and cross-sectional variations of securities would affect the impact of a speed bump, which can be tested around IEX's implementation of their speed bump. In addition, the model shows how fundamental

investors, such as large institutional investors' routing decision changes because of a speed bump. With the help of labeled transaction level data (e.g. Van Kervel and Menkveld [2018], van Kervel et al. [2018]) it can be tested as well.

References

Eric M Aldrich and Daniel Friedman. A theoretical model of the investors exchange. 2017.

James J Angel, Lawrence E Harris, and Chester S Spatt. Equity trading in the 21st century: An update. *Quarterly Journal of Finance*, 5(01):1550002, 2015.

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. High-frequency trading and price discovery. *Review of Financial Studies*, 27(8):2267–2306, 2014.

Alasdair Brown and Fuyu Yang. Slowing down fast traders: Evidence from the Betfair speed bump. 2016.

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

Haoming Chen, Sean Foley, Michael Goldstein, and Thomas Ruf. The value of a millisecond: Harnessing information in fast, fragmented markets. 2017.

Maryam Farboodi and Laura Veldkamp. Long run growth of financial technology. Technical report, National Bureau of Economic Research, 2017.

F Douglas Foster and S Viswanathan. Strategic trading when agents forecast the forecasts of others. *Journal of Finance*, 51(4):1437–1478, 1996.

Craig W Holden and Avanidhar Subrahmanyam. Long-lived private information and imperfect competition. *Journal of Finance*, 47(1):247–270, 1992.

Edwin Hu. Intentional access delays, market quality, and price discovery: Evidence from IEX becoming an exchange. 2018.

Praveen Kumar and Duane J Seppi. Information and index arbitrage. *Journal of Business*, pages 481–509, 1994.

Albert S Kyle. Continuous auctions and insider trading. *Econometrica*, pages 1315–1335, 1985.

Tomy Lee. Latency in fragmented markets. 2017.

Wei Li. High frequency trading with speed hierarchies. 2014.

Vicente Madrigal. Non-fundamental speculation. *Journal of Finance*, 51(2):553–578, 1996.

Vincent Van Kervel and Albert J Menkveld. High-frequency trading around large institutional orders. *Journal of Finance, forthcoming*, 2018.

Vincent van Kervel, Amy Kwan, and Joakim Westerholm. Order splitting and searching for a counterparty. *Manuscript, Pontificia Universidad Católica de Chile*, 2018.

Xin Wang. Why do stock exchanges compete on speed, and how? 2018.

Brian M Weller. Efficient prices at any cost: Does algorithmic trading deter information acquisition? 2016.

Liyan Yang and Haoxiang Zhu. Back-running: Seeking and hiding fundamental information in order flows. 2016.