The Externalities of High-Frequency Trading[1]

August 7, 2013

Jiading Gai Chen Yao Mao Ye

[1] All three authors are from the Department of Finance, University of Illinois at Urbana-Champaign. Please send all correspondence to Mao Ye: University of Illinois at Urbana-Champaign, 340 Wohlers Hall, 1206 South 6th Street, Champaign, IL, 61820. Email: maoye@illinois.edu. Telephone: 217-244-0474. We thank Jim Angel, Robert Battalio, Dan Bernhardt, Jonathan Brogaard, Jeffery Brown, John Campbell, John Cochrane, Robert Frank, Slava Fos, George Gao, Paul Gao, Arie Gozluklu, Joel Hasbrouck, Frank Hathaway, Pankaj Jain, Tim Johnson, Charles Jones, Andrew Karolyi, Nolan Miller, Katya Malinova, Maureen O'Hara, Neil Pearson, Andreas Park, Josh Pollet, Gideon Saar, Jeff Smith, Duane Seppi, an anonymous reviewer for the FMA Napa Conference and seminar participants at University of Illinois, SEC, CFTC/American University, University of Memphis, and University of Toronto, the 8th Annual MARC meeting and Financial Intermediation Research Society Conference 2013 and China International Conference in Finance for their helpful suggestions. We also thank NASDAQ OMX for providing the research data. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. We thank Robert Sinkovits and Choi Dongju of San Diego Supercomputer Center and David O'Neal of Pittsburg Supercomputer Center and for their assistance with supercomputing, which was made possible through the XSEDE Extended Collaborative Support Service (ECSS) program

[2] Wall Street's Need for Trading Speed: The Nanosecond Age. The Wall Street Journal, June 14, 2011.

Abstract

When price competition is constrained by tick size, speed allocates the resources due to the time priority rule. We demonstrate three implications of competition in speed. 1) We find more high frequency liquidity provision for lower price stocks with high market cap, where the one cent tick size has a higher constraint on price competition. 2) Speed has no impact on (the price) of liquidity, because speed competition already implies that liquidity providers cannot undercut each other's price. We find that exogenous technology improvements improving speed at a one millisecond, microsecond or nanosecond level do not lead to improvements on quoted spread, effective spread, trading volume or variance ratio. However, the cancellation/execution ratio increases, short term volatility increases and market depth decreases. 3) It is relative speed that matters. We find evidence consistent with the quote stuffing hypothesis (Biais and Woolley, 2011) using NASDAQ channel assignment as identification. Competition in speed but not price leads to externalities based on the canonical definition of Laffont (2008). One possible policy solution is the deregulation of tick size or decrease the importance of time priority.

Key Words: Externality, Positional Game, High-Frequency Trading, Liquidity, Price Efficiency, Quote Stuffing, Supercomputing

1. Introduction

"High frequency trading presents a lot of interesting puzzles. The Booth faculty lunchroom has hosted some interesting discussions: 'what possible social use is it to have price discovery in a microsecond instead of a millisecond?' 'I don't know, but there's a theorem that says if it's profitable it's socially beneficial.' 'Not if there are externalities' 'Ok, where's the externality?' At which point we all agree we don't know what the heck is going on."

-John Cochrane

The professional trading field is witnessing an arms race in the speed of trading. Recently, *The Wall Street Journal* stated that trading entered the nanosecond age when Fixnetix, a London-based trading technology company, announced "it has the world's fastest trading application, a microchip that prepares a trade in 740 billionths of a second, or nanoseconds." Since "investment banks and proprietary trading firms spend millions to shave ever smaller slivers of time off their activities, ...the race for the lowest 'latency' [continues], some market participants are even talking about picoseconds — trillionths of a second."[2]

Because high-frequency traders aggressively invest in technologies that reduce latency, speed should create private benefits. A recent article in Financial Times estimates that a one millisecond advantage is worth up to $100 million in annual gains.[3] However, a central issue is whether investment in speed creates social benefit. In the literature on research and development (R&D), an externality emerges if the social benefit in R&D is not consummate with the private benefit, causing inefficient over-investments (Jones and Williams, 1998 and 2000). The

[2] Wall Street's Need for Trading Speed: The Nanosecond Age. The Wall Street Journal, June 14, 2011.

[3] Speed fails to impress long-term investors; Financial Times, September 22, 2011

empirical work on the speed of trading before the sub-millisecond era argues that speed competition has social value. Hendershott, Jones, and Menkveld (2011), Hendershott and Riordan (2009 and 2011) and Hasbrouck and Saar (2013) find that speed improves liquidity. Put differently, according to the previous literature, the enhanced speed of providing or cancelling liquidity leads to better (prices of) liquidity.

However, U.S. stock markets observe price, display and time priority.[4] If liquidity providers can still undercut each other by price, speed competition is secondary. Therefore, fierce speed competition may not facilitate price competition. On the contrary, it may simply be a consequence of failed price competition. In the standard definition of the Walrasian equilibrium and the proof of Fundamental Theorem of Welfare Economics, price is infinitely divisible but time is not; all agents are assumed to arrive at the market at the same time. The reality of the financial market, however, is exactly the opposite, where time becomes divisible at the nanosecond level but price is restricted by tick size. Suppose that zero profit (or equilibrium) bid-ask spread is 0.5 cent. Then, a bid-ask spread of 1 cent loses money, but a bid-ask spread of 1 cent results in 0.5 cent of rent. As a result, there may be more supply of liquidity in the constrained spread of 1 cent than the supply with a natural spread of 0.5 cent. Thus, speed becomes the allocation rule. This paper shows two frictions, discrete price and (almost) continuous time, as regards to the canonical Walrasian equilibrium, can generate competitions we observe in the real market.

First, competition in speed is a consequence of constrained price competition. For stocks with lower price, the tick size imposes a higher constraint on price competition, which results in

[4] Orders that offer a better price have the highest execution priority. For orders with the same price, displayed orders have priority over non-displayed orders. For orders with the same displayed status, orders arriving first have the highest priority.

higher speed competition. We double sort stocks by market cap and price, and find lower price stocks have higher high frequency liquidity provision in each size category. High frequency liquidity provision is particularly large for the largest stocks with lowest price, where the natural bid-ask spread is more likely to be constrained by tick size. The price constrained stocks tend to have a large depth at the best bid and ask. High speed implies priority in a long queue. Alternatively, some traders are willing to pay to jump ahead of the queue. This strategy can be achieved by submitting orders to markets where liquidity makers need to pay a fee. This hypothesis is supported by the following empirical evidence. EDGA and EDGX are two trading platforms offered by Direct Edge. They are almost identical except in one dimension. EDGX, like most other exchanges, pays liquidity makers and charges liquidity takers. EDGA, however, has an inverted fee structure that charges liquidity providers and pays liquidity takers. Therefore, liquidity providers can effectively bypass the minimum tick size by paying a fee at EDGA. We find that stocks with lower prices have a higher market share in EDGA. Therefore, stocks with higher price constraint have higher market share in markets where price constraint can be bypassed by liquidity maker fee. In summary, price constraint leads to two important features of current market: high frequency trading and market fragmentation.

Second, when speed competition is a consequence of price constraint, liquidity is independent of speed. Speed may determine who provides liquidity, but it will not change the price of liquidity, because different liquidity providers can no longer undercut each other by price. Exogenous technology enhancements at different stages of the trading process offers us extensive clean identifications to support this hypothesis. First, we identify an enhancement of the matching engine of the NASDAQ in April 2010 and a follow-on technology enhancement from the high frequency trader side in May 2010. These two shocks have reduced the latency of submitting and processing orders from micro to nanoseconds but they do not improve liquidity measures. Quoted spread, effective spread, trading volume and variance ratio stay at the about the same level after the shocks. However, an increase in trading speed leads to a dramatic increase in the cancellation/execution ratio from 26:1 to 32:1, increasing short-term volatility as well as a decreasing market depth. We also document technology enhancements that reduce the latency of disseminating trading data from 3 milliseconds to 1 millisecond. Interestingly, NASDAQ trading data are disseminated by six identical but independent channels based on the alphabetic order of ticker symbols. Channel 1, with all the NASDAQ stocks ticker symbols beginning with A and B, was upgraded on October 10, 2011.[5] Channel 2 through Channel 6 were upgraded on October 17, 2011. This staggered technology enhancement provides a clean test to examine the casual relation between speed and liquidity. Again, we find that the enhancement of speed does not improve liquidity.

Third, time priority implies that it is the relative speed but not the absolute speed that matters. As a consequence, the economic incentive to invest in speed should be the same as the incentive to slow down other traders. Biais and Woolley (2011) and Foucault, Pagano and Röell (2013) discuss a trading strategy called "[quote] stuffing," which involves submitting a profuse number of orders to the market to generate congestions on purpose. Quote stuffing is not only a type of externality but also a type of market manipulation according to Dodd-Frank.[6] However, quote stuffing is difficult to identify. Egginton, Van Ness, and Van Ness (2011) find that intense

[5] According to the UTP plan Quotation Data Feed Interface Specification, Version 13.0e, dated Febuary 22, 2013. Each channel has a bandwidth allocation of 29,166,666 bits per second. Channel 1 handles ticker symbols from A to B; Channel 2 handles ticker symbols from C to D; Channel 3 handles ticker symbols from E to I; Channel 4 handles ticker symbols from J to N; Channel 5 handles ticker symbols from O to R; and Channel 6 handles ticker symbols from S to Z.

[6] In the Dodd-Frank Act, Section 747 specifically prohibits "bidding or offering with the intent to cancel the bid and offer before execution." On December 14, 2011, the NYSE and NYSE ARCA proposed rule 5210, which prohibits "quotation for any security without having reasonable cause to believe that such quotation is a bona fide quotation, is not fictitious and is not published or circulated or caused to be published or circulated for any fraudulent, deceptive or manipulative purpose."

quoting activity is correlated with short-term volatility, but it is not clear whether the intense episodic spikes of quoting activity are generated through "quote stuffing" or if they are natural responses to a market with higher short-term volatility (Jones, 2013). The theoretical work of Baruch and Glosten (2013), alternatively, offers a benevolent explanation of the flickering quotes. Certainly, direct inferences of quote stuffing require observations of traders' computer codes. We do, however, show suspicious patterns in the market that warrant further investigations.

Again, the identification strategy comes from the six channels that handle stock data on alphabetic order.[7] Excessive message flow of a stock stifles the information dissemination of stocks in the same channel, but stocks that reside in a different channel are not affected. Suppose a trader intends to slow down the information dissemination for stock A, he can achieve the goal by submitting messages for stock A as well as for any stock with a ticker symbol beginning with A or B. However, message flow for stock Z will not have the same effect. As a result, abnormal co-movement of message flow for stocks in the same channel is consistent with quote stuffing. We document this co-movement based on three methodologies. The first method is based on factor regressions. The idea is analogous to the literature of international finance that examines the existence of country-specific factors after controlling for the global market co-movement.[8] In our application, the six channels in total resemble a "global market," whereas each channel represents a "country." The factor regression reveals a diagonal effect: after controlling for the

[7] Slowing down the consolidated feed will cause the latency between the direct feed and the consolidated feed and thereby latency arbitrage opportunities. Most traders in the market use a consolidated data feed. High frequency traders may subscribe faster direct feed. According to Durbin (2010), however, even the most aggressive high-frequency trader still listens to consolidated feeds. For one, no market data feed is perfect; the direct feed can sometimes lose packages. Multiple sources of data help to verify that an unusual market data tick is genuine by comparing it to a second source. Also, in some cases it is possible to receive a price change from a consolidated feed sooner than a direct feed.

[8] See Lessard (1974, 1976), Roll (1992), Heston and Rouwenhorst (1994), Griffin and Karolyi (1998), Cavaglia, Brightman and Aked (2000) and Bekaert, Hodrick, and Zhang (2009), among others.

message flow of the "global market," the message flow of a stock has an abnormal positive correlation with the message flow of other stocks in its own channel. Our second identification method is a discontinuity test. We find that the first and the last stock in a channel, the order of which is based on an alphabetic sequence, have a 4.74% abnormal correlation of message flow with its own channel but zero abnormal correlations with the adjacent channels.[9] This result is further enhanced by a diff-in-diff regression. Stocks that change ticker symbols are separated into two groups. The control group changes their ticker names but not the channel assignments. The treatment group changes ticker symbols as well as the channel assignments. We find that the correlation between the treatment group's message flow and their old channels' message flow has decreased 3% after the symbol change. The correlation between the control group's message flow and their corresponding channels' message flow has remained the same after the symbol change.

Quote stuffing and increased volatility are externalities easily identifiable. An increase in order cancellation despite steady trading volume, which implies that the size of the data increases, is another externality. More importantly, competition in speed but not the competition in price, by definition, is an externality (Laffont, 2008).

Externalities are indirect effects of consumption or production activity; that is, effects on agents other than the originator of such activity which do not work through the price system. In a private competitive economy, equilibria will not be in general Pareto optimal since they will reflect only private (direct) effects and not social (direct plus indirect) effects of economic activity.[10]

[9] For the first stock in the channel, the adjacent channel is the channel immediately before. For the last stock in a channel, the adjacent channel is the channel immediately after.
[10] New Palgrave Dictionary of Economics, 2nd Edition.

By increasing his own speed, a high-frequency trader directly harms his competitors' production set of liquidity. We show that the speed of providing liquidity does not change the price of liquidity, which implies that the competition in speed does not work through the price system. Competition working through the price system does not lead to an externality, because losses to producers are precisely offset by gains to consumers (Laffont, 2008). Competition in speed, however, creates externality does not have such an effect unless the consumer of liquidity cares directly about the difference between micro and nanoseconds. In addition, quote stuffing implies that competition in speed is a positional game, in which a trader's pay-off depends on his speed relative to other traders. Traders who generate stuffing may delay themselves, but they have the economic incentive for stuffing as long as they slow other traders to a greater extent. Recent work by Frank (2003, 2005, 2008) and Bernanke and Frank (2010) argue that positional games lead to a positional externality, where any step that improves one side's relative position naturally worsens the other's ranking.

The nature of this externality offers insight for policy making. We argue that one possible policy for high-frequency trading is to liberalize the tick size or to eliminate the time priority below a certain time frame, for example, the millisecond level. Compared to the canonical economy, this is to deregulate rather than regulate the market. If the arms race in speed is due to frictions of tick size and the excessive time priority rule, a decrease of such frictions will lead to outcomes close to the first best.

This paper contributes to the literature on the impact of algorithmic and high-frequency trading. We contrast our results with the current literature that uses second or millisecond level data, which finds that high-frequency trading improves liquidity and price efficiency (Chaboud, Chiquoine, Hjalmarsson, and Vega, 2009; Hendershott and Riordan, 2009, 2011; Brogaard, 2011 a and b; Hasbrouck and Saar, 2011; and Hendershott, Jones, and Menkveld, 2011). The theoretical work on the speed of trading by Biais, Foucault, and Moinas (2011) and Jovanovic and Menkveld (2010) find that the improvement in trading speed can either increase or decrease social welfare. More specifically, Pagnotta and Philippon (2012) argue that whether innovations in speed increase or decrease social welfare depend on the initial level of speed. Allowing venues to compete on speed improves welfare if the default speed is relatively low (e.g., purely human-based trading) but decreases welfare once the default speed reaches a certain threshold. Our empirical results confirm this economic intuition: we cast doubt on the social value of increasing speed from micro to nano or pico seconds. The literature cannot assess the value of nanosecond trading due to two constraints: identification and computation.[11] We address the identification issue based on exogenous technology shocks and NASDAQ channel assignments. These identification strategies are implemented by two supercomputers from the National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE) program. To our knowledge, our empirical investigation is one of largest computing efforts ever conducted in academic finance.

This paper is organized as follows. Section 2 describes institutional details and the data. Section 3 shows the evidence of price constraint based on double sorting. Section 4 shows the impact of speed on liquidity and market efficiency measure based on exogenous technology shocks. Section 5 tests the quote stuffing hypothesis using NASDAQ channel assignment as identification. Section 6 concludes the paper, discusses the externalities generated by speed competition and the policy implications.

[11] A joint report by the Securities and Exchange Commission (SEC) and the U.S. Commodity Futures Trading Commission (CFTC) of the Flash Crash illustrates the difficulty of constructing two hours of data.

2. Data and Institutional Details

This paper uses four main datasets: NASDAQ TotalView-ITCH with nanosecond time stamp, daily TAQ data with millisecond time stamp, a NASDAQ dataset that identifies whether liquidity maker/taker is a high-frequency trader and CRSP.

2.1 Data

NASDAQ TotalView-ITCH is a series of messages that describe orders added to, removed from and executed on the NASDAQ. It is used to identify two technology shocks at the matching engine that decrease the latency from microseconds to nanoseconds. We also use ITCH data to construct a limit order book with nanosecond resolution, which is the foundation to calculate liquidity and price efficiency measures. The data come as a daily binary file and the first step is to separate order instructions into different types. The messages come with a timestamp measured in nanoseconds (10^{-9} seconds). Table 1 presents a sample of 7 types of main messages from the daily file of May 24, 2010. Messages A and F include the new orders accepted by the NASDAQ system and added to the displayable book. Message U means that the previous order is deleted and replaced with a new order. Message X provides quantity information when an order is partially cancelled. An E message is generated when an order in the book is executed in whole or in part. If the order is executed at a price that is different from the original order, a C message is generated and the new price is demonstrated in the price field. To save space, some order instructions, such as order deletion, do not indicate the stock symbol but only the reference number of the order to be deleted. It is essential to fill in the redundant details to group the order instructions based on ticker symbol, which is the foundation for the construction of the limit order book. Detailed information on how to link these messages is in the appendix. The construction of the limit order book is very data intensive. For example, we find that a message can be deleted and replaced 69,204 times using Message U. A discussion on efficient construction of the limit order book using supercomputers can be found in Gai, Choi, O'Neal, Ye, and Sinkovits (2013).

Insert Table 1 About Here

Daily TAQ data provides trades and quotes for all issues traded on the NYSE, Nasdaq (OTC), and Regionals. It has more detailed information than the usual monthly TAQ data, which is widely used in academic research. A detailed comparison between daily TAQ and monthly TAQ is in Holden and Jacobsen (2013). The unique feature that facilitates our study is the millisecond time stamp, which helps to identify another technology shock that increases the speed of disseminating trading data from 3 milliseconds to 1 millisecond. This technology improvement was adopted in a staggered matter based on the alphabetic order of ticker symbols, providing a clean identification to examine the causal effect of speed on market quality. Also, the consolidated trade file provides information on trade execution among different exchanges, which is used to calculate the market share of different trading mechanisms. This market share provides us with a nice robustness check for one of our main results. We argue that speed competition is more intense in markets with price constraint such as the NASDAQ. Price constraint leads to oversupply of liquidity and a long queue in the best ask and bid. Therefore, speed enables faster liquidity suppliers to be at the beginning of the queue with a better chance of execution. An alternative way is to supply liquidity in markets where price constraint can be bypassed. For example, EDGA, one of the two exchanges owned by Direct Edge, offers an inverted fee structure where liquidity providers need to pay a fee for executed orders. The other exchange, EDGX, has a normal fee structure where liquidity providers get the rebate while liquidity makers pay the fee. Because these two exchanges are almost identical except for the fee structure, we use the ratio of EDGA volume to EDGX as another measure of price constraint. The ratio is computed using TAQ.

NASDAQ high-frequency data are very similar to daily TAQ in structure, but it contains identifiers to distinguish high-frequency traders from the rest. This provides us with a measure for the activity of high-frequency traders. With the help of NASDAQ, we collect all three samples of datasets with high-frequency identification. These include a sample of 120 stocks selected by Hendershott and Riordan from 2008 to 2009 and February 22, 2010 to February 26, 2010. NASDAQ also provides us with the data for the same sample of 120 stocks in October 2010.

2.2. Identification of Technology Shocks

We identify three structural breaks in speed in the data and these breaks are further verified by NASDAQ and an anonymous firm. ITCH data measures time at the machining engine, which reveals two structural breaks in the speed to submit or process orders. Daily TAQ measures the time of the six independent channels that disseminate stock trading data, which provide a structural break in the speed of transmitting the data after matching. Interestingly, both of these structural changes happened on weekends, which enables the stock exchange and traders to test the new technology.

Figure 1 demonstrates the increase of speed at the matching engine level using ITCH data. Panel A demonstrates the minimum timestamp difference between two consecutive messages across the day. These two messages do not need to come from the same trader. For example, it can be the time difference between one trader's execution and another trader's cancellation. The figure shows that there is a decrease from about 950 nanoseconds to 800 nanoseconds between April 9, 2010 and April 12, 2010 and a dramatic decrease from 800 nanoseconds to 200 nanoseconds between May 21, 2010 and May 24, 2010. Panel B of Figure 1 demonstrates, for each day, the quickest execution and cancellation. As the ITCH data track the life of each individual order, we know the cancellation and execution are from the same trader. Panel B shows that the level of the fastest cancellation and execution does not change much for the April structural break, although the volatility of the fastest cancellation and execution drastically decreases. The structural break in May, however, has a dramatic impact on latency. The fastest cancellation and execution time difference decreases from about 1.2 microseconds to between 500 and 600 nanoseconds and stays below one microsecond for all but seven days after the break. Undoubtedly, NASDAQ entered the realm of nanosecond trading after May 24, 2010. Our conversion with NASDAQ reveals that the first structural break is a consequence of the installment of the Nehalem machine engine, while the second break is more likely to originate from the high-frequency trader side.

Insert Figure 1 About Here

The outflow messages on NASDAQ-listed stocks are distributed and processed across six different channels in "unlisted trading privileges" (UTP). During our sample period, Channel 1 handles ticker symbols from A to B; Channel 2 handles ticker symbols from C to D; Channel 3 handles ticker symbols from E to I; Channel 4 handles ticker symbols from J to N; Channel 5 handles ticker symbols from O to R; and Channel 6 handles ticker symbols from S to Z. The channel assignments provide us with two clean identifications. First, because the channel assignment is unrelated to firm fundamental, an abnormal correlation for message flow for stocks in the same channel is consistent with the quote stuffing hypothesis. Second, the identified technology enhancement is adopted in a staggered way with Channel 1 being the first one upgraded. This provides us a clean identification on the causal effect of technology enhancement on market quality. Figure 2 demonstrates these staggered technology enhancements. For illustration purposes, we present a 100 millisecond snapshot of the market. Basically, before Friday, October 7, 2011, the last digits of messages in daily TAQ data for all six channels end with 0, 3 or 7. This implies that information was broadcasted in either 3 milliseconds for time between 0 and 3 and 7 and 0 or 4 milliseconds for time between 3 and 7. Therefore, Panel A of figure 2 shows a minimum latency of 3 milliseconds of all six channels. Panel B shows that on Monday, October 10, 2011, Channel 1 was enhanced and was able to broadcast information in every millisecond. However, the other five channels were still broadcasting with 3 to 4 millisecond gaps. The same pattern continues to October 14, 2011, but on October 17, 2011, all six channels are able to broadcast information every millisecond.

Insert Figure 2 About Here

The data for the 100 milliseconds snapshot is representative for the whole trading hours from 9:30:00 to 16:00:00. Figure 3 demonstrates the median time gap between two broadcasts for six different channels. Before October 7, 2011, the median time gap between 2 broadcasts for all six channels is 3 milliseconds. This is a direct consequence of the fact that all the broadcasts are on milliseconds ending with 0, 3, and 7. On October 10, 2011, the median gap is 1 millisecond for Channel 1 but 3 milliseconds for channel 2-6. On October 17, 2011, all the six channels have a median gap of 1 millisecond.

Insert Figure 3 About Here

Our two identification strategy has some limitations. ITCH data feed is direct, fast data feed provided by NASDAQ. The feed, however, is single thread and we do not have cross-sectional variation in technology enhancement. To address this issue, we follow the methodology of Boehmer, Saar, and Yu (2005) by using fixed effect and control for variables known for affecting market quality. Daily TAQ data provides with cleaner identification because of cross-sectional variation, but the data is from a relatively slow consolidated feed. However, most traders in the market use a consolidated data feed. High-frequency traders may subscribe to a faster direct feed. According to Durbin (2010), however, even the most aggressive high-frequency trader still listens to consolidated feeds. For one, no market data feed is perfect; the direct feed can sometimes lose packages. Multiple sources of data help verify that an unusual market data tick is genuine by comparing it to a second source. Also, in some cases, it is possible to receive a price change from a consolidated feed sooner than a direct feed. Quote stuffing that slows down the consolidated feed will cause latency between the direct feed and the consolidated feed and thereby latency arbitrage opportunities.

3. Price Constraint, High Frequency Activity and Market Share of Different Venues

In this session, we demonstrate the following two results. More price constraint leads to 1) more speed competition in liquidity provision in exchanges with tight price constraint and 2) high market share in trading venues where price constraint can weakened by paying a liquidity maker fee. These two results also provide the economic foundation for the next section. If speed competition is a consequence of failed price competition, an increase in speed cannot change the cost of liquidity. [12]

Because stocks with a price above one dollar all have a tick size of one cent, price constraint is stricter with lower-priced stocks, especially for stocks that also have a large market cap. This intuition was suggested by The Wall Street Journal:

[12] Certainly, it decides who can provide liquidity.

"'The lower the share price, the more attractive the stock is to high-frequency traders,' said Justin Schack, managing director at Rosenblatt Securities Inc., a brokerage and stock-market research firm. 'And if you can find a stock with a low share price that is also a large-cap stock with a big float, it becomes even more attractive.'"

Our formal analysis confirms this intuition, and we interpret our results as a consequence of price constraint.

In the United States, a stock can be traded in more than 40 venues (O'Hara and Ye, 2011). One cent tick size for stocks priced above one dollar is enforced across stock exchanges. Therefore, speed competition becomes important when price competition is constrained. However, price constraint can be bypassed. While most exchanges like NASDAQ offer rebate to liquidity providers, three exchanges, Boston, BATS-Y and EDGA, offer invert fee structure that charges liquidity providers. A first glimpse would suggest that markets with inverted fees should not have any liquidity providers. Our key concept in this paper, price constraint, can explain the co-existence of these two markets. With price constraint, there is an oversupply of liquidity in markets with a maker rebate. Some liquidity providers would avoid the long queue in price-constrained exchanges by paying a fee.

We find that stocks with lower prices have more high-frequency liquidity provision as well as more market share in markets where price constraint can be weakened. The sorting criteria is almost exogenous: while price level and market cap can affect high-frequency liquidity provision and market share across different venues, it is hard to environ that high-frequency activity and market share across different venues can affect market cap.

The result is based on the data from October 2010 because it is the sample period where all measures in the section can be computed. NASDAQ high-frequency data provides the market share of high-frequency liquidity provision for 120 stocks. We use market share of EDGA and EDGX as another measure of price constraint. Because EDGA and EDGX are almost identical except the maker/taker fee, a ratio of EDGA and EDGX provides an ideal proxy for price constraint. However, EDGA and EDGX are reported on the TAQ data separately after July, 2010. Therefore, our final sample has 120 stocks from October 2010.

We sort the 120 stocks first based on market cap and then based on price. Panel A of Table 2 presents the high-frequency market making activity for these 3 by 3 portfolios. The high-frequency activity is defined as the sum of liquidity making high-frequency volume divided by total volume ((HH+NH)/(HH+NH+HN+NN)). The table reveals several facts. First, high-frequency trading market making is surprisingly low in small stocks. High-frequency traders only provide 18% to 19% of liquidity, leaving 82% to 81% liquidity provision to the remaining market participants. This pattern can be explained by lower trading volume of these stocks, but it is also consistent with price constraint. Small, illiquid stocks tend to have a higher spread and the chance to have price constraint is lower. Therefore, the competition in speed is less important. More convincing results are shown on the effect of price. For medium and large stocks, high-frequency trading liquidity provision is much higher for stocks with a lower price. For the largest stocks, high-frequency traders provide 30.9% of liquidity, whereas the number for low price stocks is 45.4%. Certainly, competition in speed is more important for the largest stocks with the lowest price, because these stocks are more likely to have price constraint.

Insert Table 2 About Here

When price is constrained and the queue for providing liquidity is long, it is easy to envision that some traders are willing to pay a small fee to jump ahead of the queue if they do not have high speed. During our sample period, three exchanges: Boston Stock Exchange, EDGA Exchange and BATS-Y offered an inverted fee structure. The volume for Boston Stock Exchange and EDGA Exchange can be identified through the TAQ data.

Panel B demonstrates the result on market share of the Boston Stock Exchange and EDGA to the consolidated volume of all trading platforms. Obviously, market share of the Boston Stock Exchange and EDGA is a decreasing function of price. For example, for the largest stocks, Boston Stock Exchange and EDGA have a market share of 8.3% for stocks with the lowest price, 5.2% for medium-priced stocks and 3.4% for stocks with the highest price. The same pattern holds for medium stocks. For example, the market share is 6.8% for medium-cap stocks with the lowest price and 3.7% for small-cap stocks with the highest price.

Certainly, traders may be attracted by the Boston Stock Exchange and EDGA for factors other than the maker/taker fee structure. To control for other heterogeneity across exchanges, we construct another variable: the ratio of volume in EDGA to EDGX. These two trading platforms are almost identical except the fee structure, which provides a cleaner test for the impact of price constraint. Panel C demonstrates the ratio of EDGA to EDGX. Again, we find that for each market cap, the market share of EDGA is a decreasing function of price. For the largest market cap, the volume of EDGA is 1.5 times the volume of EDGX for stocks with the lowest price, but the ratio is only 70% for the stocks with the highest price. For the smallest stocks, the market share of EDGA is 1.06%, 0.51% and 0.43% for low, medium and high price stocks.

4. Speed and Market Quality

Section 3 finds that speed competition may be a consequence of constrained price competition, which provides some intuition to understand the results in this section. When speed competition is a consequence of failed price competition, a reduction of latency cannot decrease the cost of liquidity. Speed may determine which trader provides liquidity, but different liquidity providers cannot undercut each other by price. Speed may provide a mechanism to allocate the rents created through price constraint, but it does not change the level of liquidity. The result is demonstrated using two tests. Such a division is based both on the types of each technology shock and the econometric method. Section 4.1 introduces our measures of cancellation, liquidity and market efficiency, which serves as the dependent variable in section 4.2 and 4.3. Section 4.2 demonstrates the effect of two technology shocks that reduce the latency of matching orders from microseconds to nanoseconds. Because the NASDAQ matching engine is single threaded, we have a single time series difference before and after treatment. We use the method by Boehmer, Saar, and Yu (2005) to control other variables that may affect market quality. Section 4.3 demonstrates the effect of the technology shock that reduces the latency of trading data dissemination from 3 milliseconds to 1 millisecond. This technology was first implemented in the channel that handles stocks with a ticker symbol beginning with A and B. This creates a double difference estimator for our analysis. The methodology of this part follows Hendershott, Jones, and Menkveld (2011). It is to be noticed that our technology shock is more exogenous than the one used by Hendershott, Jones, and Menkveld (2011). In Hendershott, Jones, and Menkveld (2011), the technology is enforced in a stagger matter, but the authors acknowledge that they are unsure where the sequence of enhancement is exogenous. Stocks with higher liquidity may be improved first. In our sample, selection rule is based on name, which provides a cleaner test to study the effect of technology enhancement on market quality. These two types of technology shocks reveal the same effect: an increase in speed at or below the 1 millisecond level does not affect our usual measure of liquidity and may worsen our usual measure of market efficiency.

4.1 Measure of Dependent Variables

Our measures of liquidity and price efficiency come from the ITCH data. We construct a message-by-message limit order book where the book is updated whenever there is a new message. That is, any order addition, execution or cancellation leads to a new order book. For example, Microsoft has about 1.08 million messages on an average trading day, and we generate and store all the resulting 1.08 million order books. This provides the most accurate view of the limit order book at any point in time. The construction is implemented by the Gordon Supercomputer in San Diego Supercomputing Center.

The message-by-message order book enables us to compute a number of metrics for market quality. We calculate four measures of liquidity. Two are spread measures: the time-weighted quoted spread and the size-weighted effective spread. The other two are depth measures: the depth at the best bid and ask and the depth within 10 cents of the best bid and ask.[13] Since we construct a full limit order book, the quoted spread is measured as the difference between the best bid and ask at any time. Each quoted spread is weighted based on the life of the quoted spread to obtain the daily time-weighted quoted spread for each stock per day. The effective spread for a buy is defined as twice the difference between the trade price and the midpoint of the best bid and ask price. The effective spread for a sell is defined as twice the difference between the midpoints of the best bid and ask and the trade price. Size-weighted effective spread is defined as the size-weighted effective spread of all the trades for each stock and each day. The two depth measures, the depth at the best bid and ask and the depth within 10 cents of the best bid and ask, are weighted using the time for each stock per day.

[13] The 10 cent cutoff is used by Hasbrouck and Saar (2011).

We also calculate two measures of price efficiency. We take the one-minute snapshot for the limit order book and calculate the minute-by-minute return based on the midpoint of the limit order book. We then measure volatility as the standard deviation of the one-minute return. We also conduct a variance ratio for price efficiency at the one-minute level. Following Lo and MacKinlay (1988), the variance ratio is defined as the variance of a two-minute return divided by two one-minute returns. In an efficient market, prices should approximate a random walk with no positive or negative correlation. Therefore, a ratio closer to 1 implies higher price efficiency.

Finally, message flow of a stock is measured by adding all types of information related to a particular stock. Cancellation ratio is defined as follows:

$$Cancellation\_execution = \frac{D+X+U}{E+C} \tag{1}$$

Basically, it measures the number of cancellations relative to executions. Finally, the daily volume measure comes from CRSP.

4.2 Shocks at Matching Engine Level: Single Time Series Difference

The technology shocks identified through the ITCH data have a single time series. We follow the methodology of Boehmer, Saar, and Yu (2005) by running regressions on the event dummy and control variables. Define L_{it} as the liquidity measure such as time-weighted quoted spread, size-weighted effective spread, time-weighted depth at the best bid and ask and time-weighted depth within 10 cents of the best bid and ask. We regress L_{it} on the event dummy and a number of controls.

$$L_{it} = \mu_i + \alpha After_t + \beta_1 logvol_{it} + \beta_2 range_{it} + \beta_3 Prc_{it} + \varepsilon_{it,}$$

(2)

$log\ vol_{it}$ is the log of the daily volume for stock *i* at day *t*. $range_{it}$ controls for volatility for stock *i* at day *t*, which is equal to the day high minus the day low in the CRSP data. Prc_{it} is the price level of the stock and μ_i is the stock fixed effect. We want to examine whether α, the coefficient for the event dummy, is significant after we control for volume, volatility and price level.

Table 3 shows that these technology shocks do not have a statistically and economically significant impact on spread. The quoted spread decreases by -0.0394 cent and the effective spread increases by 0.00115 cent, but both results are not statistically significant. The depths at the best bid and ask also do not change, but we find a 2015-share decrease of market depth within 10 cents away from the best bid and ask. Overall, we find that these two technology shocks neither increase nor decrease spread but slightly decrease the depth.

Insert Table 3 About Here

The fact that speed does not decrease spread has two natural explanations. First, the exchange follows price time priority. The competition to provide liquidity is first at price level. Time priority has a secondary role only after the price. The fact that there are intensive competitions in speed implies that there is little room for competition for price at the best bid and ask. As a result, spread might barely decrease while speed increases noticeably. Second, one argument that speed may increase liquidity is that traders with high speed can maintain a tighter bid-ask spread because they can quickly update the stale quotes before other traders can adversely select them. This argument, however, confirms that only relative speed matters: the trader with the highest speed may be able to post the tightest quotes. If the speed of all the traders increases twice, the equilibrium level of spread may not change at all. If the fastest trader is surpassed by the second fastest trader, the latter may have the ability to quote the tightest spread, but the level of spread may be the same as the original. To summarize, intensive competition in speed implies that there may be little room for further improvement in the best bid and offer. Traders with the highest speed may be able to maintain the best bid and ask spread, but the level of bid and ask is unlikely to change. We also find that market depth slightly decreases, probably because it is more risky to expose a large position when speed is higher.

For market efficiency, we follow Boehmer, Saar, and Yu (2005) and compare the mean of the volatility and variance ratio before and after the shocks without control variables. We also add the trading volume into this regression to see whether there is an increase in trading volume after these two technology shocks.

$$E_{it} = \mu_i + \lambda After_t + \varepsilon_{it,} \tag{3}$$

Therefore, we run the fixed effect regression with the dummy variable equal to 1 after the shocks. E_{it} is the price efficiency measure such as one-minute volatility and two minute to one-minute variance ratio and market volume. The variable of interest is λ, which measures the impact of these two exogenous technology improvements.

Insert Table 4 About Here

Table 4 shows that the variance ratio at the one-minute level does not have a statistically significant change before and after the technology shocks. The change in trading volume is also not statistically significant. However, volatility slightly increases after the technology shocks. The result on volatility provides additional intuition why spread does not decrease relative to the argument based on option theory. One argument that speed may improve liquidity is as follows. Limit order is the free trade option offered by liquidity providers to liquidity demanders. An increase of speed decreases duration of the option. Therefore, the price of the option decreases and liquidity improves. This argument is correct if we hold volatility as given, though the economic magnitude of the effect is not clear. When an increase of speed also increases volatility, the argument may not true. For example, in classical Black-Scholes-Merton model, the value of option depends on the product of variance and time. An increase of speed from millisecond to nanosecond may decrease the duration of option, but it also increase the volatility. An intuitive way to understand the issue is that volatility at nanosecond should be close to zero in human world, but it can be significant when multiple computers can trade in nanoseconds. Basically, speed in can allows high frequency traders cancel their trades in nanoseconds, but it can also increases volatility because it increases possible price updates in nanoseconds. Therefore, we cannot get a conclusive answer between the relationship between speed and liquidity through option pricing models, because volatility is also a function of speed. Our empirical results suggest that the impact of speed on liquidity is insignificant.

4.2 Shock that Increase Dissemination of Data: Diff-in-diff test

The six channel assignment for NASDAQ stocks provides both cross-sectional and time series variation of the technology improvements. The six channels all have 3-millisecond latency before October 7, 2011. The technology enhancement during that weekend decreased the latency to 1 millisecond for one channel while the other five channels were unaffected. Those five channels were enhanced during the next weekend. On October 17, all channels registered a 1 millisecond latency.

Those two enhancements enable us to run two diff-in-diff regressions. In the first regression, Channel 1 is the treatment group, whereas Channels 2-6 make up the control group. The before period is five trading days from October 3, 2011 to October 7, 2011 and the after period is five trading days from October 10, 2011 to October 15, 2011. In the second regression, stocks in Channels 2-6 are the treatment group, whereas stocks in Channel 1 are in the control group. The before period is from October 10, 2011 to October 15, 2011 and the after period is from October 17, 2011 to October 21, 2011. We follow the approach of and Hendershott, Jones, and Menkveld (2011) and run the following regression for liquidity measures:

$$L_{it} = \mu_i + \beta_1 pilot_i + \beta_2 after_t + \beta_3 pilot_i after_t * + \varepsilon_{it} \tag{4}$$

μ_i is the stock random effect. L_{it} is the daily liquidity measure such as time-weighted quoted spread, size-weighted effective spread, time-weighted depth at the best bid and ask and time-weighted depth within 10 cents of the best bid and ask. The variable of interest is the diff-in-diff estimator β_3. The result is summarized in Table 5. The first two columns are the effect of the first technology enhancement, and the third and fourth columns summarize the effect of the second technology enhancement. We find similar effects in the previous section: these technology shocks do not improve liquidity.

Insert Table 5 About Here

5. Test for Quote Stuffing

We have demonstrated the speed competition may be a consequence of constrained price competition and speed improvement does not improve market quality. These results question the social value of speed improvement at the sub-millisecond level. However, speed enhancement certainly has private benefit. Interestingly, time priority rule implies that it is relative speed but not absolute speed that matters. As a result, the economic incentive to slow down a competitor or an exchange should be the same as to enhance one's own speed. Biais and Woolley (2011) point out the possibility of quote stuffing, a strategy to submit an unwieldy number of orders to the market to generate congestion. As speed leads to profit, it would be equally profitable to slow down your competitors, the exchanges or both. According to Brogaard (2011c), the speed differences caused by quote stuffing are only microseconds or milliseconds, but it is enough time for a high-frequency trader to gain an advantage. The traders who generate stuffing may also delay themselves, but they still have the economic incentive for stuffing as long as it slows other traders more. This is generally the case because the generators of stuffing do not need to analyze the data they generate and they know exactly when the stuffing will occur. The other possibility raised by Brogaard (2011c) is that a malevolent trader may attempt to slow down an entire exchange. If the trader can extend the time delay between how fast an exchange can update quotes, post trades and report data, then the trader will have more time to capitalize on cross-exchange price differences. This kind of stuffing is more harmful than the previous one because it might adversely cause the breakdown of inter-market linkages, leading to sharp price movements (Madhavan, 2011). Foucault, Pagano and Röell (2013) demonstrate the benefit of quote stuffing. A flurry of order cancellations on a platform can delay the speed at which it reports trade information. In turn, this delay distorts the information sent to other participants and may create arbitrage opportunities for investors who are aware of the delay.

Quote stuffing is certainly an externality-generating activity, acting like noise or pollution in the financial market. Moreover, regulators classify stuffing as a type of market manipulation.[14] However, quote stuffing is extremely hard to identify. The seminal theoretical work of Baruch and Glosten (2013) point out that quickly cancelled orders may not be a consequence of "quote stuffing." Instead, it may simply be benevolent mixed strategy equilibrium where liquidity providers randomize the quotes. This theory provides a nice benchmark for us to test the possibility of quote stuffing.

Baruch and Glosten (2013) imply the following alternative hypothesis. Suppose flickering quotes are generated by mixed strategy, we expect that message flow of different stocks should be uncorrelated. We call it randomization hypothesis. We also have a stronger version of this hypothesis called information hypothesis, in which we believe that informationally correlated stocks should have correlated message flow. In this hypothesis, different stocks can have correlated message flow, but the correlation is driven by information.

NASDAQ channel assignment, again, provides us a clean way to test quote stuffing hypothesis against randomization hypothesis and information hypothesis. Note that quote stuffing the UTP feed may not even be the most efficient way of quote stuffing. We focus on quote stuffing the distribution of the UTP data because the channel assignment provides us with the identification strategy. Suppose, for example, a trader has information for stock A. One way

[14] In the Dodd-Frank Act, Section 747 specifically prohibits "bidding or offering with the intent to cancel the bid and offer before execution." On December 14, 2011, the NYSE and NYSE ARCA proposed rule 5210, which prohibits "quotation for any security without having reasonable cause to believe that such quotation is a bona fide quotation, is not fictitious and is not published or circulated or caused to be published or circulated for any fraudulent, deceptive or manipulative purpose."

he can delay the data distribution, and thereby the trading of stock A, is to send messages only to stock A. However, this strategy involves thousands of messages per second for one particular stock, which increases the likelihood of detection by exchanges and regulators. One way to avoid detection is to send messages to multiple tickers. A stock has an asymmetric relationship between stocks in the same channel and stocks in a different channel. For example, sending messages to ticker B will delay the trading for ticker A but sending messages to ticker Z will minutely impact stock A. It is because stock A is in the same channel as stock B but not stock Z. Therefore, we test quote stuffing based on abnormal correlations of message flows for tickers in the same channel. The abnormal correlation is not consistent with randomization hypothesis, which suggests zero correlation. It is hard to explain with the information hypothesis. If the information event is idiosyncratic, it should only affect one stock but not others. If the information event is market wide, it affects all channels equally. Unless there is a channel level information event, the abnormal correlation cannot be explained. We understand that a direct verification of quote stuffing needs data with trader identification or even the computer codes. Our results, however, do suggest suspicious behavior that deserves further investigation.

The abnormal co-movement is first demonstrated in section 5.1 using factor regression following the literature in international finance. The literature on country factor examines whether there is a country specific factor after controlling for the global market co-movement. In our setup, each channel is a "country" and all six channels are the "global market." Section 5.2 strengthens the result using a discontinuity test. We select stocks at the boundary of a channel based on alphabetic order. Next, we find that these stock have an abnormal higher correlation with stocks in the same channel than the correlation with the channel immediately next to it. Section 5.3 adopts a different-in-diff test. Stocks changing ticker symbol are divided into two groups: the pilot group, which changes ticker symbol as well as the channel, and the control group, which changes ticker symbol but remains in the same channel. We find the abnormal correlation disappears when a stock switch ticker symbol and channel but remains when stock remain in the same channel.

5.1 Factor Regression

The testing strategy follows the literature on international stock market co-movement by Lessard (1974, 1976), Roll (1992), Heston and Rouwenhorst (1994), Griffin and Karolyi (1998), Cavaglia, Brightman, and Aked (2000) and Bekaert, Hodrick, and Zhang (2009). This literature examines whether there is a country specific factor after controlling for the global market co-movement. In our context, we consider each channel as a "country" and all six channels as the "global market." We find evidence of a "channel" factor, that is, message flows for stocks in the same channel co-move with each other. This co-movement is consistent with "quote stuffing."

We divide each trading day into one-minute intervals and count the number of messages in each interval for all 2,377 stocks in the 55 trading days between March 19, 2010 and June 7, 2010. For each stock *i*, the channel message flow is the sum of all messages for stocks in channel *j* minus the message flow of stock *i*, if stock *i* is in channel *j*.[15] The market message flow is the sum of the messages for all stocks.[16] For each stock *i*, we run the following two-stage regressions following Bekaert, Hodrick, and Zhang (2009)[17]. We first regress the total number of messages of channel *j* on the market message flow:

[15] This adjustment avoids mechanical upward bias for higher correlation within the same channel.

[16] We also compute the market message flow as the sum of message flows for all stocks except stock *i*. The result is similar.

[17] As is discussed in Bekaert, Hodrick, and Zhang (2009), the first stage of orthogonalization does not change the results but only simplifies the interpretation of the coefficients. We can simply run the second stage regression and get the same result.

$$channel_{j,t} = \alpha_j + \beta_j * marketmessage_t + \varepsilon_{j,t} \quad (5)$$

The residual of this regression is defined as a new variable: $residualchannel_{j,t}$. The second step involves the following six regressions for each stock *i*:

$$f_{i,t} = \alpha_{i,j} + \beta_{i,j} * marketmessage_t + \gamma_{i,j} * residualchannel_{j,t} + \varepsilon_{i,j,t} \quad (6)$$

where $f_{i,t}$ stands for the number of messages for stock *i* at time *t*. A significant positive $\gamma_{i,j}$ when when stock *i* belongs to channel *j* is consistent with the channel effect. We also run the regression for stock *i* on the other five channels as a falsification test. Due to the large number of stocks, the coefficient of individual regression is not presented.[18] Table 6 presents the cross-sectional average of $\gamma_{i,j}$. The k^{th} column and the j^{th} row represent the average of the $\gamma_{i,j}$ if stock *i* in channel *k* is regressed on the residual message flow of channel *j*. The T-statistic is based on the hypothesis that these cross-sectional averages are zero. For example, Channel 1 has 356 stocks beginning with A and B. The first row and the second column in Table 6 demonstrate that the average coefficients of the 356 regressions of these Channel 1 stocks on the residual message flow in Channel 2 is -0.00115. The t-statistics are based on the null hypothesis that these 356 coefficients are zero.

Insert Table 6 About Here

Table 6 shows a strong diagonal effect: all the diagonal elements in the matrix are significantly positive. This means that a stock's message flow has strong positive correlation

[18] The result is available upon request.

with the message flow for the channel after controlling for the market message flow. This co-movement does not exist off-diagonal: the coefficients are negative for message flow in different channels, and most of them are statistically significant. This suggests that there is negative abnormal co-movement in message flow for stocks of different channels.

5.2 Discontinuity Test

We also supplement our regression using a discontinuity test. We first restrict our sample to stocks with at least one message in each minute. For each of the two adjacent channels, alphabetically, we pick the last stock in the previous channel and the first stock in the next channel. In other words, for Channels 2-5, we use both the first and the last stock in the channel; for Channel 1, we use the last stock, and for Channel 6, we use the first stock.[19] Panel A of Table 7 presents the ten stocks we examine. We then compare the correlation of the message flow for each stock with its own channel and the channel immediately after (before) if the stock is the last (first) one in the channel. For each stock, we first run the following regression:

$$f_{i,t} = \alpha_i + \beta_i marketmessage_t * + \epsilon_{i,t} \tag{7}$$

where $f_{i,t}$ is the number of messages for stock *i* at time *t*, and $marketmessage_t$ is the number of messages for the entire market at time *t*. We save the residual of the regression, which is the message flow after controlling for the market. We then construct two correlation variables for each stock per day: In_correlation measures the correlation between the selected stock's order flow residual with the order flow residual for stocks in the same channel, and Out_correlation measures the correlation between the selected stock's order flow residual with the order flow

[19] The first stock in Channel 1 and the last stock in Channel 6 do not have immediate alphabetic neighbors under our specification.

residual for stocks in the adjacent channel. For example, BUCY is the last stock in Channel 1. *In_correlation* is the correlation with Channel 1, while *Out_correlation* is the correlation with Channel 2. CA is the first stock in Channel 2. In_correlation is the correlation with Channel 2, while Out_correlation is the correlation with Channel 1. Panel B of Table 7 presents the results based on 550 observations (10 stocks for 55 days). We find that Out_correlation is only 0.47% and is not statistically significant; In_correlation is about 4.64%, which is 10 times as large as Out_correlation and is statistically significant. The difference between In_correlation and Out_correlation is 4.17%, with t-statistics equal to 5.11. The results based on discontinuity also suggest abnormal correlation of message flows for stocks in the same channel.

Insert Table 7 About Here

5.3 Diff-in-diff Regression

Our final test for abnormal co-movement for message flow is based on a diff-in-diff regression. We find 55 NASDAQ stocks that switch ticker symbol from January 2010 to November 18, 2011, and we separate these stocks into two groups. The stocks in the control group change ticker symbols but remain in the same channel; the stocks in the treatment group change ticker symbol as well as the channel. The control group has 13 stocks and the treatment group has 42 stocks.

We use the correlation of the stock with the channel before switching ticker as the dependent variable. For the control group, the channel assignment before and after the ticker change is the same. If a stock switches ticker from A to Z, the channel assignment will move from 1 to 6, but we always use the correlation with Channel 1 as dependent variable. The purpose of the test is to examine whether the treatment group has a decrease of correlation in message flow with the original channel after the change in ticker symbol. For each stock, we use the 30 days before the ticker change as the before period and 30 days after the ticker change as the after period.

Insert Table 8 About Here

Table 8 shows that the treatment group has a 4% decrease in correlation with the original channel after the ticker change and result is significant at the 1% level. However, the control group does not have a statistically significant reduction in correlations in message flow with the original channel. The difference between the treatment and control group reveals the channel effect: stocks have a 3% decrease in correlations with message flow after they leave a channel.

6. Conclusion

Identification and computing power impose a strict constraint for us to understand the consequence of speed competition below the microsecond level. With two identification strategies and supporting supercomputing power, we provide the first glimpse into the world of nanosecond trading.

We find that two specific technology shocks, which exogenously increase the speed of trading from the microsecond level to the nanosecond level, lead to dramatic increases in message flow. However, the increases in message flow are due largely to increases in order cancellations without any real increases to actual trading volume. The spread does not decrease following an increase in speed and the variance ratio does not improve. However, we find evidence that market depth decreases and short-term volatility increases, probably as a consequence of more cancellations. Therefore, a fight for speed increases high-frequency order cancellation but not real high-frequency order execution. Because the function of the stock market is to provide liquidity and to facilitate trading and share of risk, our results doubt the social value of decreasing latency to nanoseconds or any further decreases. We believe that investing in trading speed above some threshold should be a zero-sum game, but players may continually invest to play. Therefore, the aggregate payoff is negative even among high-frequency traders. For low-frequency traders, the externality is even more obvious. An increase in speed increases order cancellations, which generates more noise to the message flow. Low-frequency traders then subsidize the high-frequency traders because only executed trades are charged a fee. We also find a decrease of market depth and an increase of short-term volatility after the technology shocks. These findings are consistent with the observations from the market on the accumulative effects of a series of enhancements in speed. The U.S. Securities and Exchanges Commission (2010) reveals that the average trade size has decreased from 724 shares in 2005 to 268 shares as a consequence of the decrease in market depth. The increase in short-term volatility can be demonstrated by the recent plan of "Limit Up-Limit Down" to dampen volatility.

We find that stocks randomly grouped into the same channel have an abnormal correlation in message flow, which is consistent with the quote stuffing hypothesis. If the message flows of stocks are driven by market-wide information, they should affect stocks in all channels. If these message flows are driven by stock-specific information, they should be independent across different stocks. The abnormal correlation for stocks in the same channel implies that there is a "channel-level shock," which is consistent with the quote stuffing hypothesis. Since the message flow of a stock delays the trading of stocks in the same channel, but not stocks in other channels, the message flows in the same channel are more likely to co-move.

Since competition on speed is a positional arms race among high-frequency traders that creates externalities to non-high frequency traders, it is important to discuss possible solutions to this inefficiency. Currently, there are several proposed policies such as a minimum quote life, cancellation fee and transaction tax. Our paper does not provide direct evidence for these policies. Instead, our empirical evidence is consistent with a world where time priority dominates when the price competition is constrained. Several possible solutions naturally emerge.

One solution to this problem is to decrease tick size, which will force competition to focus more on price. Interestingly, from an economics point of view, some solutions would act as deregulation instead of regulation, because the current one cent tick size for stocks with a price above one dollar is imposed by regulation. The other solution is to decrease the importance of time priority below the millisecond level, where orders that arrive at the same millisecond share priority. This policy slows down the market, but it is different from minimum quote life. In minimum quote life, the constraint only applies to liquidity providers. This policy will change the relative speed of liquidity providers and liquidity demanders, which leads to an ambiguous effect. Decreasing the important of time priority, however, affects all traders in a similar manner. In summary, reducing tick size and decrease time priority will actually bring the market closer to the one described by the Walrasian equilibrium.

In the positional arms race of speed, investment tends to be mutually offsetting: suppose one high-frequency trader invests to increase the speed from micro to nanoseconds, other high-frequency traders have a strong incentive to follow. When all traders have nanosecond technology, the pay-off would not be different from the case where all traders are in microseconds. Collectively, the high-frequency traders may be better off by not investing in speed, but the individual rationale of each trader provides a strong incentive to deviate. The private solution to this problem is called the positional arms control agreement (Bernanke and Frank, 2012), in which market participants agree not to engage in mutually offsetting investments or activities. One challenge to this solution is the difficulty for a trader to verify the actions of his competitors. As a result, the consolidated audit trail to be created by the SEC is the first step for this type of solution.

A Pigovian tax can also help to correct this externality. The tax can be imposed on any investments in speed (Biais, Foucault, Moinas, 2011). Cabral (2000) discusses the tax on entry when there is a business stealth effect. The other alternative is to tax rapid order cancellation, which is accomplished through a cancellation fee. Friederich and Payne (2013) examine the impact of cancellation fee in Italian market. Also, when a trader's investment in speed can be neutralized by the same investment by his competitors in a positional game, a restriction on this type of investment may benefit all traders in the market as long as the restriction does not change the relative ranking of speed.[20] For example, on March 29, 2012, a 300 million dollar project was announced to build a transatlantic cable to reduce the current transmission time from 64.8 milliseconds to 59.6 milliseconds. According to the project's financier, "that extra five milliseconds could be worth millions every time they hit the button."[21] However, the cable may simply lead to a wealth transfer from non-subscribers to subscribers. Individual rationale makes certain high frequency traders in the transatlantic market subscribe to the cable, but when all high frequency traders subscribe to the cable, the private benefit disappears. Traders may be better off if none of them invests in the cable. Unfortunately, this cannot be sustained as equilibrium due to the private incentive to deviate. As a result, a restriction on trading speed can only be imposed by an outside authority, which can benefit all traders.

[20] In this sense, our paper does not provide a direct answer to minimum quote life policy, because minimum quote life increases the speed of execution relative to cancellation.

[21] Stock Trading Is About to Get 5.2 Milliseconds Faster. Businessweek, March 29, 2012

REFERENCES

Aghion, Philippe and Peter Howitt, 1992, A model of growth through creative destruction, *Econometrica* 60, 323-351.

Ashenfelter, Orley and David Bloom, 1993, Lawyers as agents of the devil in a prisoner's dilemma game, NBER Working paper, Princeton University and Columbia University.

Bekaert, Geert, Robert J. Hodrick and Xiaoyan Zhang, 2009, International stock return comovements, *Journal of Finance* 64, 2591-2626.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2013, Equilibrium High-Frequency Trading, Working paper, Toulouse School of Economics, HEC Paris and Toulouse School of Economics.

Biais, Bruno, and Paul Woolley, 2011, High-frequency trading, Working paper, Toulouse University and London School of Economics.

Boehmer, Ekkehart, Gideon Saar and Lei Yu, 2005. Lifting the veil: An analysis of pre-trade transparency at the NYSE, *Journal of Finance* 60, 783-815.

Brogaard, Jonathan, 2011a, The activity of high-frequency traders, Working paper, University of Washington.

Brogaard, Jonathan, 2011b, High-frequency trading and volatility, Working paper, University of Washington.

Brogaard, Jonathan, 2011c, High frequency trading, information, and profits, Working paper, University of Washington.

Brogaard, Jonathan, Terrence Hendershott and Ryan Riordan, 2013, High frequency trading and price discovery, Working paper, University of Washington, University of California, Berkley, and University of Ontario Institute of Technology.

Cavaglia, Stefano, Brightman Chris, and Michael Aked, 2000, The increasing importance of industry factors, *Financial Analyst Journal* 56, 41-54.

Chaboud, Alain, Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega, 2009, Rise of the machines: Algorithmic trading in the foreign exchange market, Working paper, Board of Governors of the Federal Reserve System.

Egginton, Jared, Bonnie F. Van Ness and Robert A. Van Ness, 2011, Quote stuffing, Working paper, Louisiana Tech University, University of Mississippi and University of Mississippi.

Foucault, Thierry, Marco Pagano and Alisa Röell, 2013, Market Liquidity, Oxford University Press.

Frank, Robert, 2003, Are positional externalities different from other externalities? Working paper, Cornell University.

Frank, Robert, 2005, Positional externalities cause large and preventable welfare losses, *American Economic Review* 95, 137-141.

Frank, Robert, 2008, Should public policy respond to positional externalities? *Journal of Public Economics* 92, 1777-1786.

Frank, Robert, and Ben Bernanke, 2010, *Principles of Economics*, McGraw-Hill.

Friederich, Sylvain and Richard Payne, 2013, Order-to-trade ratios and market quality, Working paper

Glode, Vincent, Richard C. Green and Richard Lowery, 2012, Financial expertise as an arms race, *Journal of Finance* 67, 1723-1759.

Griffin, John. M. and George Andrew Karolyi, 1998, Another look at the role of the industrial structure of markets for international diversification strategies, *Journal of Financial Economics* 50, 351-373.

Grossman, Gene M. and Elhanan Helpman, 1991, Quality ladders in the theory of growth, *Review of Economics Studies* 58, 43-61.

Hasbrouck, Joel, and Gideon Saar, 2009, Technology and liquidity provision: the blurring of traditional definitions, *Journal of Financial Markets* 12, 143-172.

Hasbrouck, Joel, and Gideon Saar, 2013, Low-latency trading, Working paper, *Journal of Financial Markets*, Available online 22 May 2013, ISSN 1386-4181

Hendershott, Terrence, and Ryan Riordan, 2011, Algorithmic trading and information, Working paper, University of Washington and University of California, Berkeley.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2011, Does algorithmic trading improve liquidity? *Journal of Finance* 66, 1-33.

Heston, Steven L., and K. Geert Rouwenhorst, 1994, Does industrial structure explain the benefits of international diversification? *Journal of Financial Economics* 36, 3-27.

Hirshleifer, J.,1971. The private and social value of information and the reward to inventive activity, *American Economic Review* 61, 561-574.

Hirschey, Nicholas H., 2012, Do High-Frequency Traders Anticipate Buying and Selling Pressure? Working paper, London Business School.

Jones, Charles I., and John C. Williams, 2000, Too much of a good thing? The economics of investment in R&D. *Journal of Economic Growth* 5, 65-85.

Jones, Charles, 2013, What Do We Know about High-Frequency Trading? Working Paper, Columbia University

Jovanovic, Boyan, and Albert J. Menkveld, 2011, Middlemen in limit order markets, Working paper, VU University Amsterdam.

Kavajecz, Kenneth A., 1999, A specialist's quoted depth and the limit order book, *Journal of Finance* 54, 747-771.

King, Benjamin F., 1966, Market and Industry factors in stock price behavior, *Journal of Business* 39, 139-190.

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tuzun Tuzun, 2011, The flash crash: The impact of high-frequency trading on an electronic market, Working paper, MIT Sloan School of Management, University of Maryland, University of North Carolina at Chapel Hill, and Federal Reserve Board.

Lessard, Donald, 1974, World, national, and industry factors in equity returns, *Journal of Finance* 29, 379-391.

Lessard, Donald, 1976, World, country, and industry relationships in equity returns: implications for risk reduction through international diversification. *Financial Analysts Journal* 32, 32-38.

Livingston, Miles, 1977, Industry movements of common stocks. *Journal of Finance* 32, 861-874.

Lo, Andrew W., A. Craig MacKinlay, 1988, Stock market prices do not follow random walks: Evidence from a simple specification test, Review of Financial Studies 1, 41-66.

Madhavan, Ananth, 2011, Exchange-Traded Funds, Market Structure and the Flash Crash Working paper

Myers, Stephen L., 1973, A re-examination of market and industry factors in stock price behavior, *Journal of Finance* 28, 695-705.

O'Hara, Maureen, Chen Yao, and Mao Ye, 2011. What's not there: The odd-lot bias in market data, Working paper, Cornell University, University of Illinois at Urbana-Champaign, and University of Illinois at Urbana-Champaign.

Patterson, Scott, 2012, *Crown Business*, New York.

Roll, Richard., 1992, Industrial structure and the comparative behavior of international stock market indices, *Journal of Finance* 47, 3-41.

Pagnotta, Emiliano and Thomas Philippon, 2012, Competing on speed, Working paper, New York University.

Tirole, J. 1988. *The theory of industrial organization*, MIT press.

Table 1: The Seven Types of Messages Used to Construct the Limit Order Book

This table provides the format of the seven types of messages used to construct the limit order book. The sample is from May 24, 2010.

Message Type	Timestamp (nanoseconds)	Order Reference Number	Buy/Sell	Shares	Stock	Price	Original Order Reference Number	Match Number	Market Participant ID
A	53435.759668667	335531633	S	300	EWA	19.5			
F	40607.031257842	168914198	B	100	NOK	9.38			UBSS
U	53520.367102587	336529765		300		19.45	335531633		
E	53676.740300677	336529765		76				7344037	
C	57603.003717685	625843333		100		32.25		20015557	
X	53676.638521222	336529765		100					
D	53676.740851701	336529765							

Table 2: High Frequency Trading and Shares of Market with Liquidity Maker Fee

Panel A demonstrates the market share of high-frequency liquidity provision. Panel B demonstrates the volume of Direct Edge A (EDGA) and Boston Stocks Exchange to the consolidated volume of all markets. Panel C is based on the ratio of Direct Edge A (EDGA) volume to Direct Edge X (EDGX). Direct Edge A and Boston Stock Exchange are markets where the liquidity maker pays the fee. The 120 stocks in NASDAQ high-frequency data are first sorted on market cap and then sorted on price.

Panel A: Percentage of Liquidity Made by High-Frequency Traders			
	Low Price	Medium Price	High Price
Small Cap	18.6%	18.9%	18.5%
Medium Cap	35.5%	23.4%	22.2%
Large Cap	45.4%	37.5%	30.9%

Panel B: Market Share of Direct Edge A and Boston Exchange			
	Low Price	Medium Price	High Price
Small Cap	3.3%	2.3%	1.8%
Medium Cap	6.8%	4.4%	3.7%
Large Cap	8.3%	5.2%	3.4%

Panel C: Direct Edge A Volume / Direct EdgeX Volume			
	Low Price	Medium Price	High Price
Small Cap	1.062	0.5135	0.4303
Medium Cap	1.9216	1.2382	1.1263
Large Cap	1.5027	1.0378	0.7028

Table 3: Speed Improvement on Submitting or Processing Orders and Liquidity

The table presents the event study of the technology shocks for the four liquidity measures. For each stock per day, *QuotedSpread* is the time-weighted quoted spread, *EffectiveSpread* is the trade size-weighted effective spread, *Depth* is the depth at the best bid and ask, *Depth10* is the cumulative depth for orders 10 cents below the best bid and 10 cents above the best ask, *after* is a dummy variable, *logvol* is the log of the daily volume, *price* is the daily price level of the stock and range equals to highest trading price minus the lowest trading price on each day for each stock. Standard errors are in parentheses, and ***, ** and * represent significance at the 1%, 5% and 10% levels, respectively.

	(1)	(2)	(3)	(4)
VARIABLES	QuotedSpread	EffectiveSpread	Depth	Depth10
after	-0.000394	0.0000115	-68.31	-2,015***
	(0.001)	(0.0003)	(93.46)	(736.50)
logvol	-0.00418***	-0.000713**	-114.6	-5,317***
	(0.001)	(0.0004)	(111.30)	(877.20)
prc	0.000907***	0.000234***	25.42**	118.3
	(0.0001)	(0.00003)	(10.66)	(83.98)
range	0.0167***	0.00441***	126.90**	-1,057**
	(0.0008)	(0.0002)	(59.91)	(472.10)
Constant	0.0596***	0.0127**	5,001***	118,697***
	(0.021)	(0.005)	(1,590)	(12,527)
Observations	5,858	5,858	5,858	5,858
R-squared	0.077	0.092	0.003	0.012
Number of ticker	118	118	118	118

Table 4: Speed Improvement on Submitting or Processing Orders, Efficiency and Volume

The table presents the event study of the technology shocks on price efficiency and volume. For each stock per day, *volatility* is the one-minute volatility, *VarianceRatio* is the one-minute variance ratio and *Volume* is the daily volume. Standard errors are in parentheses, and ***, ** and * represent significance at the 1%, 5% and 10% levels, respectively.

VARIABLES	(1) Volatility	(2) VarianceRatio	(3) Volume
after	0.0000249*	-0.00289	131,609
	(0.00001)	(0.003)	(142,487)
Constant	0.00114***	0.951***	5.971e + 06***
	(0.00001)	(0.002)	(100,625)
Observations	5,858	5,856	5,860
R-squared	0.001	0.001	0.001
Number of ticker	118	118	118

Table 5: Speed Improvement on Consolidated Tape and Liquidity

This table demonstrates the impact of speed enhancement of the consolidated tape on liquidity using diff-in-diff regression. Columns 1 and 2 demonstrate the effect of upgrading Channel 1. Stocks in Channel 1 are in the pilot group whereas stocks in Channel 2-6 are in the control group. Variable After is equal to 0 for October 3 to October 7 and equal to 1 for October 10 to October 14. Columns 3 and 4 demonstrate the effect of upgrading Channels 2-6. Stocks in Channel 1 are in the control group whereas stocks in Channel 2-6 are in the pilot group. Variable After is equal to 0 for October 10-October 14 and equal to 1 for October 17-October 21. Standard errors are in parentheses, and ***, ** and * represent significance at the 1%, 5% and 10% levels, respectively.

	Upgrade of Channel 1		Upgrade of Channel 2 to Channel 6	
VARIABLES	Effective Spread	Quoted Spread	Effective Spread	Quoted Spread
pilot	-0.000282	0.0154	0.000493	-0.0116
	(0.003)	(0.015)	(0.003)	(0.015)
after	-0.00193*	-0.00165	-0.00432***	-0.0144***
	(0.001)	(0.003)	(0.0005)	(0.001)
pilot*after	0.00167	-0.00438	-0.000227	-0.00402
	(0.001)	(0.003)	(0.0014)	(0.003)
Constant	0.0315***	0.130***	0.0355***	0.159***
	(0.002)	(0.014)	(0.001)	(0.006)

Table 6: Channel Factor Regression

This table presents the summary of the results on channel factor regression. For each stock in Channel *i*, we run six regressions:

$$f_{i,t} = \alpha_{i,j} + \beta_{ij} * marketmessage_t + \gamma_{i,j} * residualchannel_{jt} + \varepsilon_{i,j,t},$$

where *i* denotes the stock label, represents one of the six channel indices of the NASDAQ. stands for the number of the message flow for each stock at time *t*. is the message flow for all NASDAQ-listed stocks at time *t*, is the residual for regressing message flow of channel *j* on the market message flow. We run six regressions for each of the 2,377 stocks. A cell in k^{th} column and the j^{th} row in the table presents the average of the regression coefficient for those stocks belonging to channel *k* on residuals of channel *j*. Therefore, the diagonal elements present the stock's co-movement with the same channel, while the off-diagonal elements present the stock's co-movement with a different channel. The t-statistics for the hypothesis are in the parentheses. ***, **, * represent the statistical significance at the 1%, 5% and 10% levels, respectively.

Dependent Variable / Independent Variable	Channel 1 Message Flow	Channel 2 Message Flow	Channel 3 Message Flow	Channel 4 Message Flow	Channel 5 Message Flow	Channel 6 Message Flow
Channel 1 Residual	0.00304**	-0.00115**	-0.00079*	-0.00087*	-0.00082***	-0.00105*
	(2.267)	(-2.132)	(-1.696)	(-1.848)	(-3.049)	(-1.753)
Channel 2 Residual	-0.00049***	0.00300***	-0.00017	-0.00034***	-0.00032***	-0.00028
	(-6.219)	(4.340)	(-1.532)	(-2.425)	(-2.768)	(-1.480)
Channel 3 Residual	-0.00039***	-0.00020*	0.00209***	-0.00043***	-0.00052***	-0.00045**
	(-4.810)	(-1.708)	(5.553)	(-2.687)	(-3.005)	(-1.962)
Channel 4 Residual	-.00049***	-0.00045**	-0.00049**	0.00266***	-0.00054***	-0.00031
	(-3.979)	(-2.092)	(-2.256)	(3.869)	(-2.348)	-1.297
Channel 5 Residual	-0.00074**	-0.00068	-0.00094*	-0.00085***	0.00310*	-0.00072
	(-2.273)	(-1.492)	(-1.868)	(-3.869)	(1.738)	(-1.158)
Channel 6 Residual	-.00042***	-0.00026**	-0.00036***	-0.00022***	-0.00032***	0.00186***
	(-8.172)	(-2.191)	(-3.448)	(-2.790)	(-4.794)	(6.227)

Table 7: Discontinuity Test

This table presents the results from the discontinuity test. Panel A lists stocks used for the discontinuity test: based on the alphabetical order, they are the first and last stock in each channel with a minimum of one message per minute. In_correlation measures the correlation between the selected stock's order flow residual with the order flow residual for stocks in the same channel, and Out_correlation measures the correlation between the selected stock's order flow residual with the order flow residual for stocks in the immediately adjacent channel. Panel B presents the results based on 550 observations (10 stocks for 55 days).

Panel A		
	In_correlation	Out_correlation
BUCY (Last in Channel 1)	Correlation between BUCY and Channel 1 stocks	Correlation between BUCY and Channel 2 stocks
CA (First in Channel 2)	Correlation between CA and Channel 2 stocks	Correlation between CA and Channel 1 stocks
DWA (Last in Channel 2)	Correlation between DWA and Channel 2 stocks	Correlation between DWA and Channel 3 stocks
EBAY (First in Channel 3)	Correlation between EBAY and Channel 3 stocks	Correlation between EBAY and Channel 2 stocks
ITRI (Last in Channel 3)	Correlation between ITRI and Channel 3 stocks	Correlation between ITRI and Channel 4 stocks
JBHT (First in Channel 4)	Correlation between JBHT and Channel 4 stocks	Correlation between JBHT and Channel 3 stocks
NWSA (Last in Channel 4)	Correlation between NWSA and Channel 4 stocks	Correlation between NWSA and Channel 5 stocks
ONNN (First in Channel 5)	Correlation between ONNN and Channel 5 stocks	Correlation between ONNN and Channel 4 stocks
RVBD (Last in Channel 5)	Correlation between RVBD and Channel 5 stocks	Correlation between RVBD and Channel 6 stocks
SAPE (First in Channel 6)	Correlation between SAPE and Channel 6 stocks	Correlation between SAPE and Channel 5 stocks

Panel B: Differences After Control for Market Message Flow			
In_correlation	Out_correlation	In_correlation-Out_correlation	t-statistics
0.0464	0.00474	0.0417***	5.11

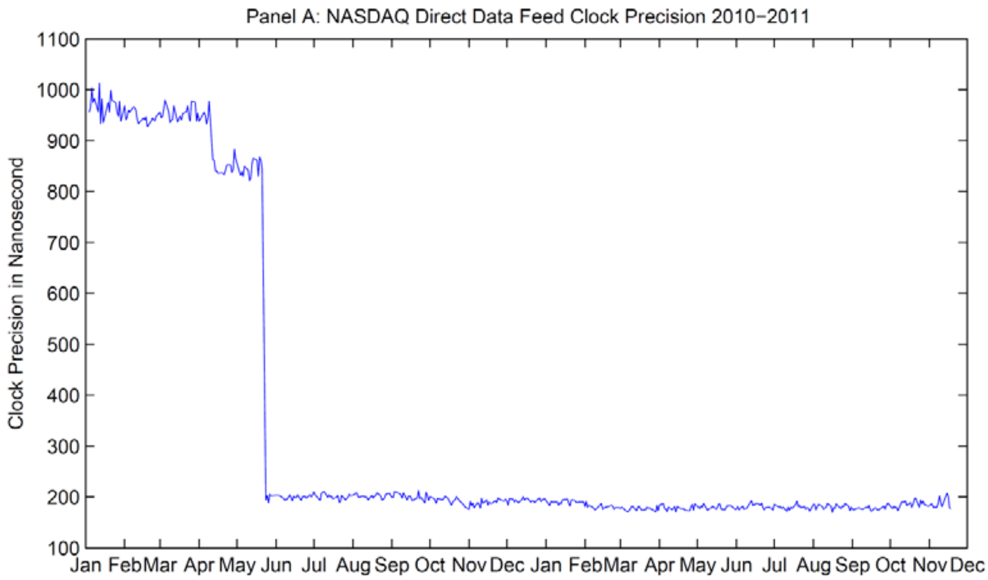
Table 8. Diff-in-diff Test

This table presents the diff-in-diff regression for 55 stocks that switch ticker symbol from January 2010 to November 18, 2011. The stocks in the control group change ticker symbol but remain in the same channel; the stocks in the treatment group change ticker symbol as well as the channel. The before period has 30 days before the ticker change and the after period has 30 days after the ticker change. The dependent variable is the message flow correlation with the original channel.

Diff-in-diff Table			
	Treatment Group	Control Group	Diff
Before	0.485***	0.507***	-0.0222**
	(0.00519)	(0.00916)	(0.0106)
After	0.444***	0.495***	-0.0513***
	(0.00523)	(0.00921)	(0.0106)
Diff	-0.0414***	-0.0123	-0.0291*
	(0.151)	(0.013)	(0.015)

Figure 1: Technology Shocks on Submitting and Cancelling Orders

These figures demonstrate the impact of our two technology shocks on latency. The first technology shock happened between April 9, 2010 and April 12, 2010. The second shock happened between May 21, 2010 and May 24, 2010. We have two measures of latency. Panel A demonstrates the minimum time differences between two consecutive messages for the NASDAQ market. Panel B demonstrates the fastest cancellation and execution for the NASDAQ market.



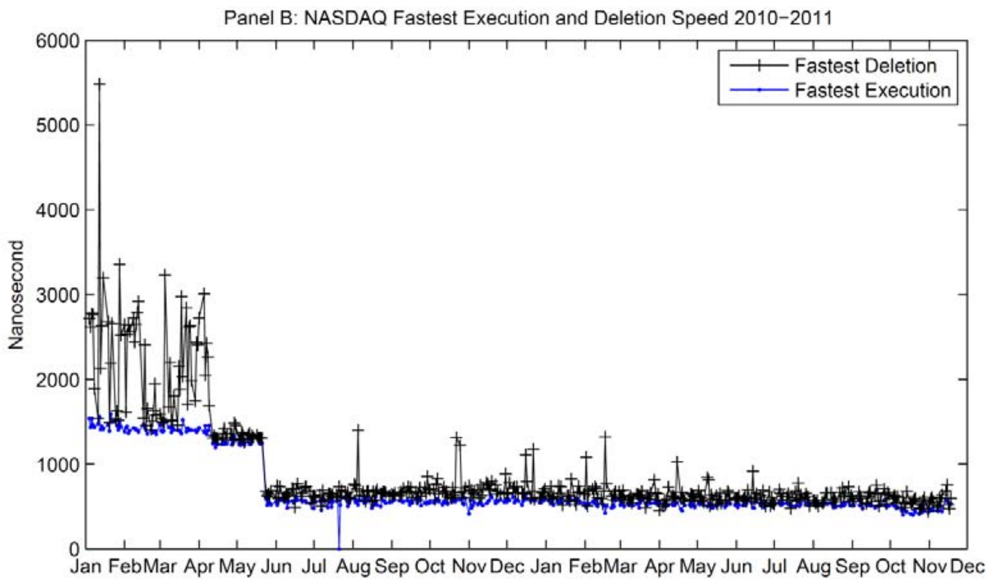


Figure 2: Technology Shocks on Transmitting Trading Data

This graph demostrates the technology shocks on the consolidated tape. Before October 7,2011 messages are broadcast only on the millisonds ending with 0, 3 or 7, implying a minimum gap of 3 milliseconds. On October 10, 2011 Channel 1 was able to breadcast information every millisecond but other channels continue to have gaps till October 14, 2011. On October 17, 2011, Channel 2-6 are also upgraded and there is no gap to broadcast messages.

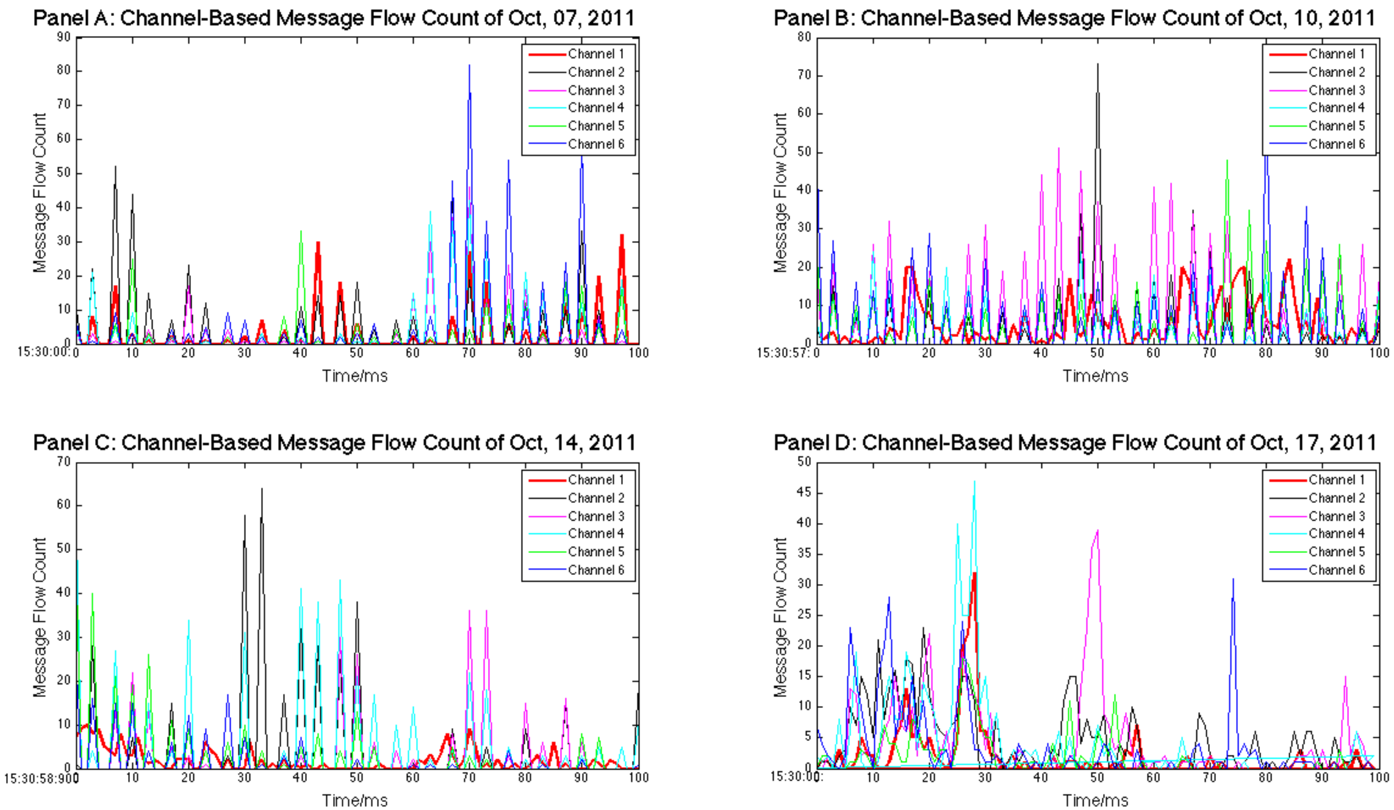
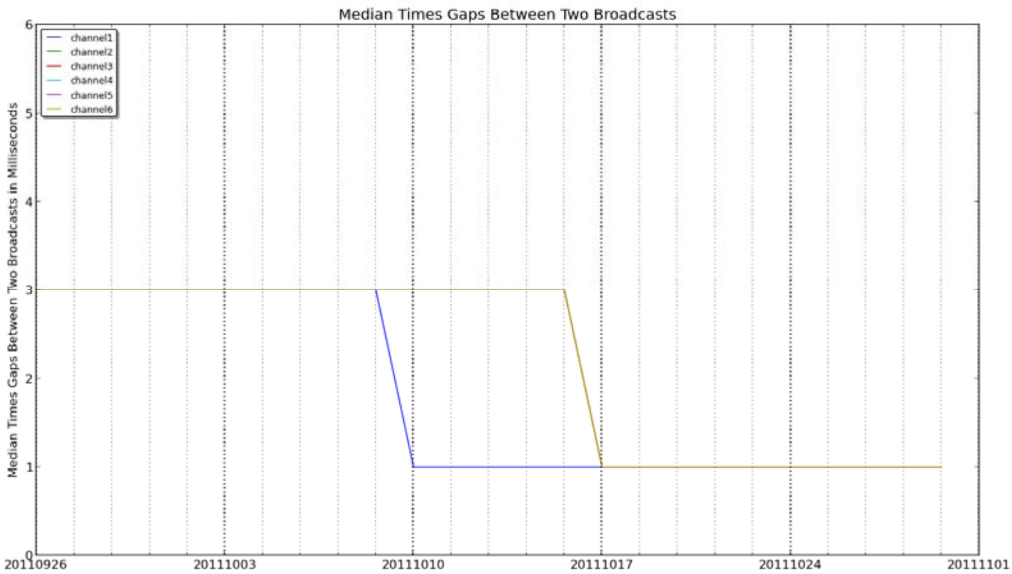


Figure 3: Median Time Gaps between Two Broadcasts

This graph demonstrates the median times gaps between two broadcasts based on all the quotes in the six channels. The horizontal axis represents the trading days between September 26, 2011 and November 1, 2011, and the vertical axis represents the time in milliseconds. Before October 7, 2011, the median time gap between 2 broadcasts for all six channels is 3 milliseconds. On October 10, 2011, the median gap is 1 millisecond for Channel 1 but 3 milliseconds for channel 2-6. On October 17, 2011, all the six channels have a median gap of 1 millisecond.



Appendix: Structure of ITCH Data

This appendix provides more information on how to construct the limit order book from messages in ITCH data. The type of information is listed in Table 1.

Messages A and F include the new orders accepted by the NASDAQ system and added to the displayable book. NASDAQ assigns each message a unique reference number. Messages A and F include the timestamp, buy/sell reference number, price, amount of shares and the stock symbol. The only difference between messages A and F is that F indicates the market participant identification associated with the entered order. The first message in Table 1 is an A message with a reference number 335531633 to sell 300 shares of EWA at $19.50 per share. Time is measured as the number of seconds past midnight. Therefore, this order is input at second 53435.759668667, or 14:50:35:759668667. The F message shows a 100-share buy order for NOK at a price of $9.38 per share with UBSS as the market participant. Message U means that the previous order is deleted and replaced with a new order. The update can be on the share price or quantity of shares. In our example, order 335531633 has a change in price from $19.50 to $19.45, generating a new order with reference number 336529765. To conserve space, message U does not indicate the ticker symbol and the buy/sell reference number. Only after the trader finds the reference number for the first time the updated message was deleted can she link the updated message back to message A or message F to locate its ticker symbol and buy/sell reference number. In our example, we can link order 336529765 to the original order 335531633 and know that it is a sell order for EWA. We find that a message can be deleted and replaced 69,204 times using message U. In short, new orders can originate from three message files: messages A, F and U.

Message X provides quantity information when an order is partially cancelled. Orders with multiple partial cancellations share the same reference number. Message X only contains a timestamp, order number and the quantity of shares cancelled. We need to link the X message to the original A or F message in order to find the stock in our sample and update its limit order book. In our example, the X instruction deletes 100 shares from order 336529765. The U message with reference number 336529765 implies that the size of the order is reduced to 200 shares at a price of $19.45 per share. However, we need to link the U message to the A message to know that new order is to sell EWA.

An E message is generated when an order in the book is executed in whole or in part. Multiple executions originated from the same order share the same reference number. An E message also only has the order reference number and the quantity of shares executed. Therefore, we need to trace the order to the original A or F message to find the stock and the buy/sell information. In our example, the order reference number first points to the U message (336529765), which then tracks to an A message. Now we know that a sell order for EWA is executed; however, the price information is from the U message, where the price has been updated from $19.50 to $19.45 per share. After matching, the system will generate a matching number of 7344037. If the order is executed at a price that is different from the original order, a C message is generated and the new price is demonstrated in the price field.

Message D provides information when an order is deleted. All remaining shares are removed from the order book once message D is sent. In our example, all the remaining shares of order 336529765 are deleted. The order originally had 300 shares, and an X message deletes 100 shares from the book, while an E message leads to an execution for a sale of 76 shares. Therefore, message D deletes 124 shares from the book. The price level is $19.45 per share, which is known from the U message, and the stock and the buy/sell indicator can be found at the A message.