The Externalities of High-Frequency Trading

Jiading Gai Chen Yao Mao Ye[1]

August 31, 2012

Abstract

We show that exogenous technology shocks that increase the speed of trading from microseconds to nanoseconds dramatically increase the order cancellation/execution ratio from 26:1 to 32:1 but do not have any detectable impact on liquidity, price efficiency or trading volume. We find evidence consistent with "quote stuffing," which involves submitting an abnormally large number of orders followed immediately by a cancellation to generate order congestion. The stock data are handled by six randomly grouped channels in NASDAQ, and the message flow of a stock can slow down the trading of stocks in the same channel but not stocks in a different channel. We detect an abnormally high level of co-movement of message flow for stocks in the same channel using factor regression and a discontinuity test. Our results suggest that an arms race in speed at the sub-millisecond level is a positional game, where a trader's pay-off depends on her speed relative to other traders. Private benefit then leads to offsetting investments on speed by different high-frequency traders even if there is no social benefit.

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

[1] All three authors are from 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, Jonathan Brogaard, John Campbell, John Cochrane, Slava Fos, Maureen O'Hara, Frank Hathaway, Tim Johnson, Charles Jones, Andrew Karolyi, Neil Pearson, Gideon Saar, and Duane Seppi for their helpful suggestions. We thank NASDAQ OMX for providing the research data and National Science Foundation's XSEDE (Extreme Science and Engineering Discovery Environment) program for their research support.

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 has witnessed an arms race in the speed of trading. Recently, *The Wall Street Journal* stated that trading entered the nanosecond age when London-based trading technology company Fixnetix announced that "it has the world's fastest trading application, a microchip that prepares a trade in 740 billionths of a second, or nanoseconds." However, "investment banks and proprietary trading firms spend millions to shave ever smaller slivers of time off their activities." "With the race for the lowest 'latency' continuing, some market participants are even talking about picoseconds — trillionths of a second."[3]

Regulators and academics across the Atlantic, however, are proposing policies to slow down the trading process. One policy is minimum quote life. SEC Chairwoman Mary Shapiro, for example, stated that she would consider a minimum "time in force" for quotations. "The likely minimum duration for a quote under such a proposal could be 50 milliseconds, which has been suggested by several sources."[4] In Europe, the *Review of the Markets in Financial Instruments* (MiFID) solicits comments on "How should

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

[4] Minimum Quote Life Faces Hurdles. *Traders Magazine*, November 15, 2010.

the minimum period be prescribed?"[5] The other recommendation is a cancellation fee. The fee would be assessed based on the average number of order cancellations to actual transactions by market participants to properly absorb the externalized cost of their activity.[6] Finally, the academic research by Biais, Foucault, and Moinas (2011) proposes Pigovian tax on investment on speed such as colocation.

Because high-frequency traders invest aggressively in reducing latency, speed should create private benefits for them. However, whether or not an increase in speed creates social benefits is not clear. If the social benefit is not consummate with the private benefit, an externality emerges and there would be an inefficiently large investment in technology. The Securities and Exchange Commission (SEC) intends to regulate orders with lives of less than 50 milliseconds (fleeting orders henceforth) with the belief that speed and the investment in speed create externalities. However, there has been no empirical evidence supporting the externality of cancellation and speed of trading. In addition, one can argue from a theoretical ground that decreasing the trading speed leads to lower liquidity. For example, traders who submit limit orders provide trading options. A longer time-in-force for the option would increase the cost of the trading option and thereby drain liquidity.

This paper examines possible externalities generated by speed of trading and cancellation. Two obstacles prevent researchers from examining these questions: identification and computation capabilities. The identification problem arises naturally due to the endogenous relationship between liquidity, price discovery, order cancellation and speed. For example, Egginton, Van Ness, and Van Ness (2011) find that intense quoting activity is correlated with short-term volatility. However, it is hard to establish a causal relationship. Even less clear is whether or not the intense episodic spikes of quoting activity are generated through manipulative "quote stuffing" or as a natural response to a market with higher short-term volatility. Computing power is also a serious challenge. A joint report by the SEC and the U.S.

[5] *European Commission Public Consultation: Review of the Markets in Financial Instruments Directive* (*MiFID*), February, 2011, page 7.
[6] Recommendations Regarding Regulatory Responses to the Market Events of May 6, 2010: Summary Report of the Joint CFTC-SEC Advisory Committee on Emerging Regulatory Issues, page 11.

Commodity Futures Trading Commission (CFTC) on the flash crash illustrates the difficulty of constructing two hours of data during the flash crash. Interestingly, while regulators in Europe and the United States try to search for the best cut-off value for minimal quote life, we have not seen simple summary statistics like the mean and median quote life. Even descriptive statistics are hard to compute in the high-frequency trading world, let alone an in-depth analysis. This paper finds two identification strategies to examine the impact of speed of trading and cancellation, and the identification strategies are implemented by two supercomputers. To our knowledge, our empirical investigation is one of largest computing efforts ever in academic finance.

We find that two consecutive technological shocks decrease latency from microseconds to nanoseconds. These shocks drastically increase both the trading speed and the cancellation ratio, which escalates from 26:1 to 32:1. However, there is no impact on trading volume, spread, depth, or price efficiency due to these shocks. The evidence shows that the increase of speed from microseconds to nanoseconds facilitates order cancellation without an impact on liquidity or price efficiency. The theoretical work on the speed of trading by Biais, Foucault, and Moinas (2011), Jovanovic and Menkveld (2010), and Pagnotta and Philippon (2012) is based on the following trade-off: on one side, high-frequency traders may detect new trading opportunities and increase social welfare; on the other side, high-frequency trading may cause an adverse selection problem for slow traders and generate externalities to them. Our result implies that the second effect dominates at the sub-millisecond level. While an increase in speed from seconds to milliseconds may result in more trading opportunities, an increase in speed from microseconds to nanoseconds may only lead to wealth transfer from slow to fast traders. In the spirit of the classical work on informal economics by Hirshleifer (1971), the distributive aspect of speed provides a motivation for investing in speed that is quite apart from — and may even exist in the absence of — any social usefulness of speed. As a result, an externality emerges.

As speed provides private value to a trader, it is almost equally valuable to slow her competitors down. Biais and Woolley (2011) discuss a trading strategy called "(quote) stuffing," which involves

submitting an unwieldy number of orders to the market to generate congestion. Stuffing is certainly a type of externality-generating behavior. Moreover, regulators classify stuffing as a type of market manipulation.[7] Our identification strategy for stuffing is based on the channel assignment of NASDAQ stocks. The trading data of NASDAQ stocks are handled by six independent channels based on the alphabetic orders of ticker symbols.[8] This channel assignment is close to random with respect to firm fundamentals, which provides us with the identification scheme for one type of quote stuffing.[9, 10] We first document clear evidence of abnormal co-movement of message flow for stocks in the same channel through factor regression. The result is further strengthened through a discontinuity test. We pick the first and last stock in the channel based on alphabetic order, and find that the stock has an abnormal correlation with its own channel rather than with an adjacent channel.[11] This result is consistent with quote stuffing, because the excessive message flow of a stock slows down the trading of the stock in the same channel, but does not have the same effect on stocks in a different channel.

We show that the cancelations now consume 97% of computer system resources that the whole market has to bear. This is similar to the externality generated through the tragedy of the commons. The

[7] 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."

[8] 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.

[9] Quote stuffing can also happen in other steps of the trading process. For example, before an order is matched, there are exchange gateways to check the validity of the orders, such as whether or not the trader has the necessary margin requirement. There are multiple gateways for an exchange. Therefore, one strategy of quote stuffing involves stuffing all the gateways except one. The trader causing the quote stuffing uses the one gateway he does not stuff, while other traders need some time to figure out which gateway is not stuffed.

[10] This type of quote stuffing affects consolidated data feed. Most traders in the market use a consolidated data feed. Some high-frequency traders may subscribe direct data for some direct data feed for some market centers while using a consolidated data feed for some other market centers. The most aggressive high-frequency trading firms will have a direct market data feed from every exchange. However, according to Durbin (2010), 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. Having multiple sources of data helps to verify that an unusual market data tick is real and not an error by having a second source to compare it to. Also, it possible in some cases to get a price change on a consolidated feed sooner than from a direct feed.

[11] 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.

high level of cancellations forces stock exchanges and all the traders to continuously upgrade trading systems and increase bandwidth to accommodate higher message flow. In addition, most stock exchanges only charges fees for executions but not cancellations. These worsen the externality problem, because traders executing trades are subsiding traders with lots of cancellations. We also show that an increase in speed increases the cancellation-to-execution ratio. We believe that it is because the increase in speed leads to more time periods for a fixed calendar time, which increases the number of possible moves for a trading game among high-frequency traders. However, the aggregated opportunity for trading or providing liquidity does not increase, which explains why we do not observe an increase in trading or liquidity. When the cancellation-to-execution ratio increases, the externality from the tragedy of the commons becomes worse.

This paper contributes to the literature on the impact of algorithmic and high-frequency trading. As the first paper to explore the impact of high-frequency trading in a nanosecond trading environment, our results contrast the literature using second or millisecond level data, which find 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 increase in trading speed from second to millisecond may still have social benefit by creating new trading opportunities, but we doubt whether such benefit exists from a reduction from microseconds to nanoseconds, and maybe in the future, to picoseconds. This further increase in trading speed has fundamentally affected the stock market. According to Patterson (2012), the average holding period of stocks was about eight months in 2000. In 2008, the holding period was two months. This number decreased to 22 seconds in 2011. While it is naïve to eliminate high-frequency traders, it is equally naïve to let the arms race of speed proceed without any restriction.

As the first paper to document the externality of high-frequency trading, our work provides rationale and the direction for policy intervention for this activity. The arms race of speed is a typical example of positional externality (Frank, 2005), where a trader's pay-off depends on her speed relative to

other traders. This externality would lead to positional arms races, which are a series of mutually offsetting investments on speed. The offsetting investment is led by a standard prisoner's dilemma problem, where individual rationale will lead to the continuation of investing in speed. Yet the combination of choices is worse for all the high-frequency traders, rather than the alternative in which each trader does not continue to nanoseconds or picoseconds trading. In this case, a speed limit for the orders may benefit all high-frequency traders. A speed limit will also benefit long-term traders who are not already in the speed game. Our results show that an increase in speed increases the cancellation ratio, which implies that a decrease of speed results in fewer cancellations, which may solve the constraint in bandwidth. Also, we find that an increase of speed from microseconds to nanoseconds does not increase liquidity or price efficiency. As a result, we conjecture that a decrease of speed from nanoseconds to microseconds does not harm liquidity or price efficiency. The analysis also provides justifications for a cancellation fee. Because of the externalities we document, speed and cancellation appear more attractive to individual traders than to society as a whole. A Pigovian tax will help to solve the problem. One way is to tax the investment in speed, as is suggested by the theoretical work of Biais, Foucault, and Moinas (2011). The other way is to tax the cancellation, which is exactly the cancellation fee.

More broadly, our paper is related to the literature of overinvestment in research and development, information acquisition, professional services, and financial expertise. All these overinvestments are generated by the divergence of the private and social benefit, or externality. Jones and Williams (2000) argue that whether there is an underinvestment or overinvestment in technology depends on whether the investment generates a positive or negative externality. Hirshleifer (1971) models two types of information: foreknowledge of states of the world that will be revealed by nature itself (e.g., earning announcements), and discovery of hidden properties of nature that can only be laid bare by action. We conjecture that information that exists at the microsecond or nanosecond level is more of foreknowledge. Traders who are the first to get the information can reap dramatic financial benefits, but the benefit for society from their trading activities is close to zero. The general notion that agents may

overinvest to compete in a zero-sum game goes back at least to Ashenfelter and Bloom (1993), who find that outcomes of labor arbitration hearings are unaffected by legal representation, as long as both parties have lawyers. A more recent work by Glode, Green, and Lowery (2011) examines the arms race for financial expertise. Frank (2005, 2008) documents the arms races in positional goods — those for which relative position matters most. Our analysis, particularly on quote stuffing, demonstrates that it is relative speed that matters in the microsecond and nanosecond trading environment.

This paper is organized as follows. Section 2 describes the data. Section 3 provides the summary statistics and preliminary results. Section 4 examines quote stuffing based on the channel assignment of NASDAQ. In Section 5, we construct two limit order books, a full limit order book with all the orders and another with orders with a quote life of 50 milliseconds or above, to see the contribution of fleeting orders to liquidity and market efficiency. In Section 6, event studies are used to compare the market quality before and after the system enhancement of speed. Section 7 concludes.

2. Data

2.1. NASDAQ TotalView-ITCH Data

We use NASDAQ TotalView-ITCH for our analyses. The data are a series of messages that describe orders added to, removed from, and executed on the NASDAQ. The data comes as a daily binary file with all the order instructions. Our first step is to separate order instructions into different types. To conserve space, we focus on seven types of messages: A, F, U, E, C, X, and D. A complete list of message types can be found in the NASDAQ TotalView-ITCH data manual. The messages come with timestamp measured in nanoseconds (10^{-9} seconds).

Table 1 presents a sample of each type of message from the daily file of May 24, 2010. The daily file contains the order instructions of all the NASDAQ stocks. 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 for each stock.

Messages A and F include the new orders being 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 or sell reference number, price, amount of shares, and the stock symbol. The only difference between the A and F types is that F indicates the market participant identifiers associated with the entered order but A does not. 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. The order was input at second 53435.759668667, or 14:50:35: 759668667 because the time is measured as the second past midnight. The F message shows a 100-share buy order for NOK at a price of $9.38/per share, and the market participant to enter the order is UBSS.

A U message means that order previously added is deleted and replaced by new messages. The update can be on the price or number of shares. In our example, order 335531633 has a change in price from $19.50 to $19.45, and a new order with reference number 336529765 is generated. To converse space, message U does not indicate the ticker symbol and the buy or 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 track its ticker symbol and buy or 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 “U” message. In short, order addition can originate from three message files, Message A, F, and U.

A message X file provides quantity information when an order is partially cancelled. Orders with multiple partial cancellations share the same reference number. Message X only has a timestamp, order number, and the number of shares cancelled. We need to link the X message to the original A or F message to find the stock in our sample that we want to update the limit order book. In our example, the

X instruction deletes 100 shares from order 336529765. The U message with 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 the operation was done 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 number of shares executed. Therefore, we need to trace the order to the original A or F message to find that stock and buy-sell information. In our example, the order reference number first points to a U message (336529765), then the U message tracks to the A message. Then we know that a sell order of 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.

A message D file 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 used to have 300 shares, and an X message deleted 100 shares from the book and an E message leads to an execution for a sale of 76 shares. Therefore, a D message deleted 124 shares from the book. The price level is $19.45/per share, which is from the U message, and the stock and buy/sell indicator can be found at the A message.

2.2. Sample Stocks and Periods

We construct two samples of stocks for our study. The test for quote stuffing uses the message flow of all the 2,377 common stocks listed in NASDAQ. The construction of message-by-message limit order books requires a large amount of computing power and storage space. Therefore, we start from the same 120 stocks selected by Hendershott and Riordan (2011a, b) for their NASDAQ high-frequency

dataset. These stocks provide a stratified sample of securities representing differing market capitalization levels and listing venues. The sample of stocks has been used by a number of recent studies, such as those by Brogaard (2011 a, and b), Hendershott and Riordan (2011a, b), and O'Hara, Yao, and Ye (2011). Because Hendershott and Riordan picked the stocks in early 2010 but our sample period extends to 2011, 118 of the 120 stocks are left in the sample.

With the help of NASDAQ, we identify two exogenous technology shocks that significantly reduce latency. Interestingly, both technology shocks occurred during weekends, a more convenient time for exchanges or traders to test their technology enhancement as the exchanges are closed. One technology shock happened between April 9, 2010 and April 12, 2010, and the other one occurred between May 21, 2010 and May 24, 2010. A discussion of these two shocks is in Section 5.1. For our event study, we select 15 trading days before and after the first shock (March 19, 2010 to April 9, 2010 and April 12, 2010 to April 30, 2010) and 10 days before and after the second shock (May 10, 2010 to May 21, 2010 and May 24, 2010 to June 7, 2010). The reason is that we want to exclude the effect of the flash crash in our before-after comparison. The tests for quote stuffing and partial equilibrium analysis do not involve the event dummy, and we find similar results by including or excluding the week of the flash crash. In summary, our results on quote stuffing and partial equilibrium analysis are based on the 55 trading days from March 19, 2010 to June 7, 2010 and our event study results are based on 50 trading days of data by excluding the week of the flash crash.

2.3. Construction of the Variables

Our test on quote stuffing is based on the time series pattern of aggregated message flow. The aggregated message flow is defined as the sum of the 7 types of NASDAQ messages. Other types of messages are mostly stock symbol directory information and administrative information, such as trading halt and trading resumption. We use the stock directory information to link the NASDAQ messages to each stock and use the administrative information when we construct the limit order book, but we do not

count the stock symbol and administrative information into the total message flow. The result is similar if we add the symbol and administrative information because there are less than 10 observations per stock per day.

The cancellation ratio can be defined in two ways. The first measure of cancellation is based on the number of entered orders. We define the cancellation ratio as 1 minus the number of trades divided by the number of entered orders, that is,

$$Cancellation\_ratio = 1 - \frac{E+C}{A+F+U}. \quad (1)$$

The second measure is based on cancelled orders. We define the cancellation and execution ratio as:

$$Cancellation\_execution = \frac{D+X+U}{E+C}. \quad (2)$$

The U type message is in both definitions because a U message involves both additions and deletions of orders. These two measures are not exactly the same because of such issues as partial cancellation or multiple executions from the same order, but certainly they are very highly correlated.

We define the order life as the difference between order entry through A, F or U message information and order deletion through D, X or U message information. We also compute the life for orders that are executed, but we focus on orders that are cancelled or updated unless otherwise indicated. The results are very similar if executed orders are included, because the number of executed orders is much less the number of cancelled updated orders.

We also use A, F, U, E, C, X, and D message to construct the limit order book with nanosecond resolution. The traditional way to construct limit order books is based on Kavajecz (1999). The idea is to construct a snapshot of limit order books on a fixed time interval such as 5 minutes or 30 minutes. We

examine the impact of fleeting orders, thus a lot of information is lost if the analysis is based on snapshots. Therefore, 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 stock has about 1.08 million messages on an average trading day, and we generate and store all the resulting order books. This provides the most accurate view of the limit order book at any point in time. In Section 5, we are also able to construct two limit order books: one with all the orders and the other one with orders with a life greater or equal to 50 milliseconds.

3. Preliminary Results

Figure 1 provides a histogram of quote life for cancelled orders with a life less than one second, with each bin in the graph representing five milliseconds. The sample includes 118 stocks for which we construct the limit order books; 30% of the observations fall into the bin with the shortest quote life. This pattern has a natural explanation: traders who send fleeting orders want to cancel their orders as fast as possible, and the limit is how quickly they can cancel them. Our preliminary evidence has two implications. First, there are some recent discussions on the optimal minimum quote life. Two of the most commonly referred proposals are 50 and 100 milliseconds. However, Figure 1 shows that whether we set the minimal quote life at 50 milliseconds or 100 milliseconds would not generate a large difference in market outcome. There are very few observations between 50 and 100 milliseconds. Throughout the paper, we use 50 milliseconds as the cut-off value, but we believe that the results based on 100 milliseconds are similar. Second, the constraint for decreasing quote life is technology, which provides us justification for using system enhancement as a shock to quote life.

Table 2 presents the order cancellation ratio and the ratio of orders with a life less than 50 milliseconds, or fleeting orders. The result is sorted by the ratio of fleeting orders. Erie Indemnity Company (Ticker ERIE) has the highest cancellation ratio, as well as the highest ratio of fleeting orders, with 99.56% of submitted orders being cancelled and 68.78% of orders being cancelled within 50

milliseconds. The rankings based on cancellation ratios are slightly different, but closely correlated. Some of the most liquid stocks have very high cancellation ratios, as well as a high percentage of fleeting orders. For example, 96.09% of orders of Apple (AAPL) are cancelled, with almost 40% cancelled within 50 milliseconds. For orders of Google (GOOG), 95.92% are cancelled, with 30% cancelled within 50 milliseconds. The high cancellation ratio means that, on average, there is only one trade for every 30 orders, while the ratio is 232 to 1 for ERIE. The median level of cancellation is 96.5%, which implies an execution ratio of 28 to 1. The mean level of fleeting orders is about 15% across stocks, with the median equal to 13%.

Table 3 demonstrates the position of fleeting orders. Hasbrouck and Saar (2009) find that most fleeting orders are placed inside best bid and offer (BBO) in 2004, which is consistent with the strategy of detecting hidden liquidity. In our sample, however, only 11.25% of the fleeting orders are placed inside BBO, while 52.23% are placed at BBO and 36.53% are placed outside the BBO,[12] which suggests that fleeting orders are placed with different purposes in 2004 and 2010.

With the help of NASDAQ and an anonymous firm, we identify two structural breaks in latency. We use these two structural breaks as an identification strategy to examine the impact of speed on market quality. Interestingly, both of these structural changes happened on weekends, which is likely because both the exchanges and the traders usually test new technology during weekends. The first structural break happened between Friday, April 9, 2010 and Monday, April 12, 2010. NASDAQ confirms it was due to the installation of the Nehalem matching engine. A more dramatic break happened between Friday, May 21, 2010 and Monday, May 24, 2010. The change is due to the change in speed from the high-frequency traders' side. These technology shocks are exogenous because they are not correlated with the level of liquidity or price discovery in the market. The private benefit to become the fastest exchange and trader is so large that it is beneficial to implement and use the innovation once it is mature.

[12] Fleeting orders are defined as order with a life less than two seconds in Hasbrouck and Saar (2009). In our sample, they are defined as orders less than 50 milliseconds.

Figure 2 shows the impact of these two technology shocks on latency. Panel A demonstrates the result on 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 message and another trader's cancellation message. The figure shows that there is a decrease from about 950 nanoseconds to 800 nanoseconds between April 9 and April 12, and a dramatic decrease from 800 nanoseconds to 200 nanoseconds from May 21, 2010 to May 24, 2010. Panel B of Figure 2 demonstrates, for each day, the quickest execution and cancellation for the day. As the ITCH data track the life of each individual order, we know the cancellation and execution is 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 decrease. The structural break in May, however, has a dramatic impact on latency. The quickest cancellation and execution decrease from about 1.2 microseconds to 500-600 nanoseconds, and stay below one microsecond for all but seven days after the break. Therefore, NASDAQ trading enters nanosecond trading regime after May 24, 2010.

4. Test for Quote Stuffing

Biais and Woolley (2011) define quote stuffing as submitting an unwieldy number of orders to the market to generate congestion. To be more specific, traders who cause stuffing can slow down other traders because 1) there are more messages in the queue for the exchange to process; 2) the dissemination of trading data from the exchange is delayed so that their competitors cannot react promptly to market conditions; and 3) their competitors need to analyze the data, but they do not. Quote stuffing is certainly an externality-generating activity, like noise or pollution in the financial market.

We believe that quote stuffing is perfectly incentive compatible in positional arms races. In the trading environment of microseconds or nanoseconds, it not the *absolute* speed, but the *relative* speed to competitors and stock exchanges that matters. As speed leads to profit, it would also be equally profitable

to slow down your competitors, the exchange or both. The economic incentives for enhancing speed and slowing down others should be the same, if it is relative speed that is important. According to Brogaard (2011 c), the speed differences caused by quote stuffing are only microseconds or milliseconds, but that is enough time for a trader to gain an advantage. The traders who generate stuffing may also slow down 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 stuffing will occur. The other possibility raised by Brogaard (2011 c) is that a malevolent trader may be trying to slow down an entire exchange. If the trader can extend the time delay between how fast an exchange can update its quotes, post trades, and reports data, then the trader will have more time to take advantage of cross-exchange price differences. This kind of stuffing is more harmful than the previous one, because it might effectively cause the breakdown of inter-market linkages, leading to sharp price movements (Madhavan, 2011).

We provide a formal test of quote stuffing based on the following identification strategy. The outflow messages on NASDAQ listed stocks are distributed and processed across six different channels in "unlisted trading privileges" (UTP).[13] The six channels have the same breakout for the UTP Quotation Data Feed (UQDF) and the UTP Trade Data Feed (UTDF). In total there are 2,377 stocks reported to UTP in our sample period. The channel assignment provides an ideal identification for quote stuffing. Note that quote stuffing the UTP feed is not the only way to accomplish quote stuffing. As is explained by footnote 8 and 9, quote stuffing may also happen at the exchange gateway or the matching engine and attacking the UTP feed may not even be the most efficient way of quote stuffing. We focus on quote stuffing the distribution of the Trade and Quote (TAQ) data because the channel assignment provides us with the identification.

Suppose, for example, a trader has information for stock A. One way he can slowdown the data

[13] Although the NASDAQ also trades stocks listed in other exchanges, the outflow messages of other exchanges is handled by different systems. Quote data from other exchanges are handled by the Consolidated Quote System (CQS) and the trade data of other exchanges is handled by the Consolidated Tape System (CQS).

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 immediately attracts the attention of the exchanges and regulators. Therefore, one way to avoid detection is to send messages to multiple tickers. Then, a stock has an asymmetric relationship with stocks in the same channel and stocks in a different channel. For example, sending messages to ticker B will slow down the trading for ticker A, but sending messages to ticker Z will have a much smaller impact on stock A. It is because A is in the same channel as stock B but not stock Z. Therefore, we test quote stuffing based on abnormal correlations of message flow for tickers in the same channel.

4.1 Factor Regression

We obtain the channel assignments for NASDAQ-listed stocks from NASDAQ. In our sample period, there are six channels for NASDAQ-listed stocks. 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; Channel 6 handles ticker symbols from S to Z. The testing strategy follows the literature on industry factors by Meyers (1963), King (1966), and Livingston (1977), and 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). The idea is that we consider each channel as a "country" and all the six channels as the "global market." The literature on country factor examines whether there is a country factor after controlling for the global market co-movement. Using the same method, we find the evidence of a "channel" factor, that is, message flow for stocks in the same channel co-moves 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 the 2,377 stocks in the 55 trading days from March 19, 2010 to June 7, 2010. For each stock i, the channel message flow is the sum of all message flows for stocks in the channel j minus the

message flow of stock i, if stock i is in channel j. We make this adjustment to avoid mechanical upward bias to find that a stock have higher correlation with message flows in its own channel. The market message flow is the sum of message flow of 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:

$$channel_{j,t} = \alpha_j + \beta_j * marketmessage_t + \varepsilon_{j,t.} \quad (3)$$

We save the residual of this regression as a new variable, $residualchannel_{j,t}$. In the second step, we run the following six regressions for each stock i:

$$f_{i,t} = \alpha_{i,j} + \beta_{i,j} * marketmessage_t + \gamma_{i,j} * residualchannel_{j,t} + \varepsilon_{i,j,t,} \quad (4)$$

where $f_{i,t}$ stands for the number of messages for stock i at time t. $\gamma_{i,j}$ measures the channel-level effect after controlling for the market-wide effect. We are particularly interested in $\gamma_{i,j}$ when stock i belongs to Channel j. However, we also run the regression for stock i on other channels as a falsification test. Due to the large number of stocks, we do not present the coefficients for individual regressions, but the results are available upon request. Table 4 provides the summary statistics of all these regressions. A cell in the k^{th} column and the j^{th} row in the table presents the average of the $\gamma_{i,j}$ coefficient if stock i in Channel k is regressed on the residual message flow of Channel j. For example, the coefficient in the first row and the second column, -0.00115, means that the average regression coefficients of the Channel 1 stocks on the residual message flow in Channel 2 is -0.00115. The t-statistics are based on the hypothesis that these coefficients are zero. The results show 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 with the

[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.

message flow for the channel even after controlling for market message flow. We also find that this type of co-movement does not exist between stocks in different channels: the coefficients are negative for message flow in different channels, and most of them are statistically significant.

4.2 Discontinuity Test

We also supplement our regression using the discontinuity test. 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 with at least one message in each minute. 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 used the first stock.[18] Panel A of Table 5 presents the ten stocks we examined. We then compare the correlation of the message flow for each stock with its own channel and the channel immediately adjacent (before) if the stock is the last (first) one in the channel. For each stock, we first run the following regression:

$$f_{i,t} = \alpha_i + \beta_i marketmessage_t * + \epsilon_{i,t,} \quad (5)$$

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 for each 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 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 5 presents the results based on 550 observations (10 stocks for 55 days). We find that *Out_correlation* is

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

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 result based on discontinuity also suggests abnormal correlation of message flow for stocks in the same channel.

5. The Contribution of Fleeting Order to Liquidity and Price Discovery

We construct message-by-message limit order book to document all the updates of the market across the day. This enables us to evaluate the liquidity contribution of the fleeting orders. For example, if an order of 100 shares improves the bid-ask spread by 1 cent for 49 milliseconds, its contribution to liquidity is one cent multiplied by 49 milliseconds. Its contribution to depth at the best bid and ask is 100 shares multiplied by 49 milliseconds. Suppose the current best ask is $20, then a new limit sell order of 200 shares at $20 does not improve the spread, but it improves the depth by 200 shares. Its weighted contribution to depth is 200 shares multiplied by the time of the improvement. Snapshots of the limit order book are also generated through the message-by-message limit order book when we compute minute-by-minute returns for the short-term volatility and variance ratios.

In this section we evaluate the contribution of fleeting orders to liquidity and price discovery by constructing two limit order books: one with all the orders and the other excluding orders with a life less than 50 milliseconds. We call this partial equilibrium analysis because we do not consider the complex dynamics if the SEC enforces the 50-millisecond minimum quote life. We supplement the partial equilibrium analysis with a natural experiment in the next section. We speculate that if the SEC enforces the minimal quote life of 50 milliseconds, current orders with a life fewer than 50 milliseconds are more likely to be at 50 milliseconds rather than completely disappear. Therefore, a comparison of a limit order book *with* all orders and a limit order book *without* limit orders with a 50-millisecond quote life or lower provides an upper bound for the contribution of fleeting orders to liquidity. We find that fleeting orders provide very little liquidity to the market and do not improve market efficiency, which is consistent with

the finding in the next section that higher speed does not improve liquidity or price efficiency.

5.1. Contribution of Fleeting Orders on Spread and Depth

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 asks and the depth within 10 cents of the best ask and bid. Because we construct the 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 and each 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 midpoint of the best bid and ask price 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 asks and the depth within 10 cents of the best ask and bid, are weighted using the time for each stock each day. [19]

The results for our four liquidity measures for the 118 stocks for 55 days from March 19, 2010 to June 7, 2010 are shown in Table 6. The daily measure for one stock is an observation. Table 6 shows that the average quoted spread for the whole book is about 5.971 cents and the median is about 2.805 cents. The effective spread is lower, with a mean of 3.63 cents and a median of 1.85 cents. The removal of some fleeting orders would increase the quoted spread because some of them improve the bid-ask spread. We find that a limit order book without fleeting orders has a quoted spread of 5.996 cents, reflecting a 0.0251 cent increase in quoted spread on average. The increase in relative terms is 0.215% of the bid-ask spread. Therefore, fleeting orders contribute to 0.215% liquidity to the market in terms of spread. The measure is much smaller based on the median spread. The fleeting orders decrease the quoted spread by 0.0000378 cents in terms of median spread, which is 0.116% of the liquidity. The result for effective spread is

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

slightly larger: a limit order book without fleeting orders has a 0.0399 cent increase in effective spread in terms of mean spread and a 0.0095 cent increase in median spread.

Fleeting orders contribute 3.96 shares of liquidity in terms of the depth at the best ask, and 3.59 shares in terms of the depth at the best bid. The number for median spread is again much lower. On a median day for a median stock, fleeting orders contribute to 0.24 shares for the depth on the ask side and 0.22 shares on the bid side. On average, fleeting orders contribute to 28 shares to the depth within 10 cents of the best bid and ask, and the median number is 0.54 shares for best ask and 0.56 shares for the best bid. In conclusion, fleeting orders do contribute to the spread and depth, but the effect is trivial. Also, our estimation of the contribution of spread and depth from fleeting orders is an upper bound. The fleeting orders would not completely disappear if a minimum quote life of 50 milliseconds is imposed. We would expect some to be exactly at 50 milliseconds if the rule is imposed.

5.2. Contribution of Fleeting Orders to Price Efficiency

While a limit order book without fleeting orders must have lower liquidity by construction, the result for volatility and price efficiency is less clear. Depending on their position, fleeting orders can either increase or decrease volatility or price discovery. Table 7 presents this these results. We take the one-minute snapshot for the limit order book with and without fleeting orders and calculate the midpoint based on the one-minute return. Panel A presents the stock date comparison, where each stock for each day is one observation. Interestingly, for 2,774 cases, or 43.08% of the stock date observations, fleeting orders have no impact on volatility based on the one-minute return because the one-minute return volatility of the full limit order book is the same as the limit order book without fleeting orders. There are 1,657 cases, or 25.73% of the observations, where the full limit order book has higher volatility than the limit order book without fleeting orders. There are 2,010 cases, or 31.21% of the observations, where the full limit order book has a lower volatility. Panel B shows that the differences in volatility, although statistically significant, are economically trivial. For example, the median volatility for the full limit order

book is 0.0010046, while the limit order book without fleeting orders has a volatility of 0.0010057. The difference is only 0.0000009. Therefore, fleeting orders decreases short-term volatility, although not by much.

We also conduct the variance ratio for price efficiency at the one-minute level, and the result is not statistically significant. 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 one implies higher price efficiency. Panel A of Table 7 shows that in 42.52% of the cases, the limit order book with fleeting orders has the same variance ratio as the limit order without fleeting orders. Panel B compares the differences in variance ratios, which are neither economically nor statistically significant. In fact, if we measure the difference between the variance ratio and one, the test based on means suggest that the return in the full limit order book is closer to a random walk, while the test on the median suggest the opposite. Both tests, however, are not statistically significant. Therefore, fleeting orders neither increase nor decrease the price efficiency at the one-minute level, the time frame that people can observe.

6. Natural Experiment

The partial equilibrium analysis provides an estimation of the contribution of fleeting orders to liquidity and price discovery, but it suffers from endogeneity issues. We supplement our partial equilibrium analysis with technology shocks that exogenously increase the speed of trading. Again, we find that these exogenous shocks do not improve liquidity or price efficiency; however, they do increase the cancellation ratio.

6.1 Identification of Technology Shocks

These two structural breaks, particularly the second, result in a dramatic increase in the cancellation ratio. For the ten event days before and after the second structural change, the mean

cancellation/execution ratio increases from 25.82 to 32.04, while the cancellation/execution ratio increases from 20.30 to 33.56 from March 2010 to June 2010.

6.2 Effects of the Technology Shock

To evaluate the effect of the technology shock, we follow the approach of Boehmer, Saar, and Yu (2005) and Hendershott, Jones, and Menkveld (2011), who run regressions on the event dummy and control variables. Suppose the liquidity measure can be written as:

$$L_{it} = \mu_i + \alpha After_t + \beta_1 vol_{it} + \beta_2 range_{it} + \beta_3 Prc_{it} + \varepsilon_{it,} \quad (3)$$

where L_{it} is the measure of liquidity such as quoted spread, effective spread, and depth; μ_i is the stock fixed effect; vol_{it} is the daily volume for each stock each day; $range_{it}$ is the measure of the volatility in terms of day high minus day low in the CRSP; and Prc_{it} is the price level of the stock. Our interest is on whether α, the coefficient for the event dummy, is significant after we control for volume, volatility, and price level.

One limitation of our identified technology shocks is that both are close to the May 6, 2010 flash crash. Therefore, we eliminate May 3 to May 7, the week of the flash crash. We then have 15 trading days before the first technology shock (March 19 to April 9) and 15 trading days after (April 12 to April 30). Then we have 10 trading days before the second technology shock (May 10 to May 21) and 10 trading days after the second technology shock (May 24 to July 7). To increase the power of the test, we group the two before and the two after the period in our test, but still find that the technology shock does not increase liquidity.

Table 8 shows that the technology shocks do not have a statistically and economically significant impact on liquidity. The coefficient for the quoted spread is only -0.000394, which is not statistically significant. The coefficient for effective spread is even smaller (0.0000115). We do find that the quoted depth within 10 cents of the best ask and bid becomes worse after the technology shock, but the quoted

depth at the best ask and bid does not change.

For market efficiency, we follow Boehmer, Saar, and Yu (2005) and compare the mean of the volatility and variance ratio before and after the shock event without a control variable. Therefore, we run the fixed effect regression with the dummy variable equal to one after the shock. Table 9 shows that the volatility slightly increases after the technology shock, although it is only significant at the 10% level, with a magnitude of only 0.0000249. The variance ratio before and after the technology shock is also not statistically different. Finally, when we examine the change of trading volume before and after the technology shock, and the change is not statistically significant.

6.3. Summary

We find that the two technology shocks have a large impact on the cancellation/execution ratio but not on volume, liquidity or price efficiency. We believe that an increase in trading speed increases the number of periods for how the trading game is played between different high-frequency traders. Therefore, we see more cancellations, probably because more complex games result in more cancellations. For example, the quote stuffing strategy may need more orders to generate congestion. However, an increase in speed does not improve liquidity or price efficiency. However, speed may create several externalities. Quote stuffing is certainly one type. Even without quote stuffing, we argue that investment in speed with sub-millisecond accuracy may provide private benefits to traders without consummate social benefit. Therefore, there may be an overinvestment in speed. Finally, the exchanges continuously makes costly system enhancements to accommodate higher message flow, but this enhancement facilitates more cancellations, not additional trading. Because the current exchange fee structure only charges trades not cancellations, traders who want to trade subsidize those who cancel, reflecting a wealth transfer from long-term traders to high-frequency traders.

7. Conclusion

Identification and computing power impose a strict constraint for us to understand the intent and consequence of high-frequency cancellation. With two identification strategies and supporting supercomputing power, we provide the first glimpse into the world of microsecond and nanosecond.

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 the channels. If these message flows are driven by stock-specific information, they should be i.i.d. 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. Because the message flow of a stock slows down the trading of stocks in the same channel but not stocks in other channels, the message flow in the same channel is more likely to co-move.

We also find that fleeting orders, or orders with a life less than 50 milliseconds, have trivial contributions to liquidity and no contributions to price efficiency. We also find that technology shocks that exogenously change the speed of trading from the microsecond level to the nanosecond level lead to a dramatic increase in message flow. However, the increase is largely an increase in cancellations without a real increase in volume.

Market liquidity does not increase with this increase in speed, nor does price efficiency. This result has two implications. First, a fight for speed increases high-frequency cancellation but not real high-frequency trading. Because the function of stock market is to provide liquidity and to facilitate trading and share of risk, our results provide a question on the social value of decreasing latency to nanoseconds or even lower latency. We believe that investing in trading speed above some threshold should be a zero-sum game, but players need to continuously 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 cancellations, which generates more noise to the message flow. Low-frequency traders then subsidize the high-frequency traders because only trades not cancellations are charged.

This paper also has following policy implications. As investment in speed is a positional arms race, there is a divergence between private and social benefit. A Pigovian tax helps to correct this externality. The tax can be imposed to the investment on speed such as colocation (Biais, Foucault, Moinas, 2011). The other alternative is to tax high speed cancellation, which is exactly the cancellation fee. Also, when a trader's investment in speed can be neutralized by the same investment of her competitors in a positional game, a restriction on this type of investment may benefit all the traders as long as the restriction does not change the relative ranking of speed among traders. [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 builder of the cable, "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 the non-subscribers to subscribers. Individual rationale makes certain high-frequency traders in transatlantic market subscribe the cable, but when all high frequency traders subscribe the cable, even the private benefit disappears. Traders may be better off if none of them invest in the cable. However, this cannot be sustained as equilibrium due to the private incentive to deviate. In this case, restriction on speed can only be imposed by an outside authority, and such a restriction 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

Ashenfelter, O., & Bloom, D. (1993). Lawyers as Agents of the Devil in a Prisoner's Dilemma Game, NBER Working Paper

Bekaert, G., Hodrick, R. J., & Zhang, X. 2009. International stock return comovements. *The Journal of Finance,* 64(6), 2591-2626.

Biais, B., Foucault, T., & Moinas, S. 2011. Equilibrium High-Frequency Trading, Working Paper.

Biais, B., and Woolley, P. 2011. High-frequency trading. Working paper, Toulouse University, IDEI.

Boehmer, E., Saar, G., & Yu, L. (2005). Lifting the veil: An analysis of Pre-trade transparency at the NYSE. *The Journal of Finance, 60*(2), 783-815.

Brogaard, J. A., 2011a. The activity of high-frequency traders. Working Paper.

Brogaard, J. A., 2011b. High-frequency trading and volatility. Working Paper.

Brogaard, J. A., 2011c. High frequency trading, information, and profits, Working paper

Cavaglia, S., C. Brightman, and M. Aked, 2000, The increasing importance of industry factors. *Financial Analyst Journal*, 41-54.

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

Egginton, J., Van Ness, B., & Van Ness, R. 2011. Quote stuffing, Working Paper.

Glode, Vincent; Green, Richard C.; and Lowery, Richard, 2011, Financial Expertise as an Arms Race, *The Journal of Finance, forthcoming.*

Griffin, John. M., & Andrew Karolyi, G. 1998. Another look at the role of the industrial structure of markets for international diversification strategies. *Journal of Financial Economics,* 50(3), 351-373.

Hasbrouck, J., and Saar, G. 2009. Technology and liquidity provision: The blurring of traditional definitions. *Journal of Financial Markets*, 12(2), 143-172.

Hasbrouck, Joel, and Gideon Saar, 2011a. Low-latency trading. Manuscript, Cornell University.

Hendershott, T. J., and Riordan, R. 2011b. High-frequency trading and price discovery. Working Paper.

Hendershott, Terrence, and Ryan Riordan. 2011b. Algorithmic trading and information, Working Paper.

Hendershott, T., Jones, C. M., & Menkveld, A. J. 2011. Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1), 1-33.

Heston, S. L., and Rouwenhorst, K. G. 1994. Does industrial structure explain the benefits of international diversification? *Journal of Financial Economics,* 36(1), 3-27.

Hirshleifer, J.,1971. The private and social value of information and the reward to inventive activity. *The American Economic Review, 61*(4), 561-574.

Hirschey, Nicholas H., 2012, Do High-Frequency Traders Anticipate Buying and Selling Pressure? Working Paper.

Jones, C. I., and Williams, J. C. 2000. Too much of a good thing? the economics of investment in R&D. *Journal of Economic Growth,* 5(1), 65-85.

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

Kavajecz, K. A. 1999. A specialist's quoted depth and the limit order book. *The Journal of Finance*, 54(2), 747-771.

King, Benjamin F., 1966. Market and Industry factors in stock price behavior. *The 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. Manuscript, U of Maryland.

Lessard, Donald. 1974. World, national, and industry factors in equity returns. *Journal of Finance*, 29(3), 379-391.

Lessard, Donald 1976. World, country, and industry relationships in equity returns: implications for risk reduction through international diversification. *Financial Analysts Journal*, 32(1), 32-38.

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

Lo, A. W., MacKinlay, A. C. 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies, 1*(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. *The Journal of Finance*, 28, 695-705.

O'Hara, Maureen, Chen Yao, and Mao Ye, 2011. What's not there: The odd-lot bias in TAQ data. Working Paper.

Patterson, Scott, 2012, Dark Pools, Crown Business, New York.

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

Pagnotta, E., Philippon, T. 2012. Competing on speed, Working Paper.

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	19.5000			
F	40607.031257842	168914198	B	100	NOK	9.3800			UBSS
U	53520.367102587	336529765		300		19.4500	335531633		
E	53676.740300677	336529765		76				7344037	
C	57603.003717685	625843333		100		32.2500		20015557	
X	53676.638521222	336529765		100					
D	53676.740851701	336529765							

Table 2: Percentage of Fleeting orders and the level of Cancellation

This table presents the percentage of orders cancelled (Cancel Ratio) and the percentage of orders cancelled within 50 milliseconds (Fleeting Percent).

Stock	Fleeting Percent	Cancel Ratio	Stock	Fleeting Percent	Cancel Ratio	Stock	Fleeting Percent	Cancel Ratio
ERIE	68.78	99.56	BRCM	15.03	93.77	ROCK	10.43	97.37
CRVL	43.43	99.49	PFE	15.01	93.86	LSTR	10.31	96.61
NC	41.58	99.57	AA	14.98	94.92	CELG	9.90	93.68
AAPL	39.51	96.10	FFIC	14.88	98.02	DCOM	9.44	97.39
GOOG	30.19	95.92	BW	14.82	98.69	MOD	9.27	96.74
PPD	28.89	99.22	BAS	14.78	95.55	CPWR	9.25	94.73
CKH	28.10	98.75	GLW	14.66	94.18	KMB	9.10	96.32
PNC	27.83	97.43	CTRN	14.60	97.80	SFG	9.06	98.32
AMZN	26.52	96.68	AMGN	14.23	92.95	FMER	9.06	94.46
LANC	25.28	98.30	HPQ	13.97	94.80	CNQR	8.89	96.35
ROG	24.60	99.42	DELL	13.94	93.68	MIG	8.79	96.47
AZZ	24.03	99.29	ABD	13.71	95.79	MMM	8.71	96.25
AXP	23.35	95.57	AMAT	13.58	93.22	CBEY	8.59	95.59
DOW	22.31	95.01	FULT	13.55	94.66	RVI	8.44	96.87
BHI	21.70	96.44	LECO	13.55	98.37	LPNT	8.28	96.30
FRED	21.67	95.95	FL	13.38	95.31	MAKO	8.27	96.42
KTII	21.51	96.82	PNY	13.34	97.60	BXS	8.24	97.47
JKHY	21.38	97.22	BRE	13.18	96.72	COO	8.07	97.58
HON	21.14	96.61	CSCO	13.06	93.91	ROC	8.01	97.40
IPAR	18.88	98.49	GENZ	12.84	93.70	MDCO	8.01	92.92
NXTM	18.81	96.56	EBAY	12.82	93.10	CSL	7.97	97.67
CMCSA	18.65	94.59	MANT	12.50	97.29	IMGN	7.82	94.07
SWN	18.50	95.10	COST	12.46	95.54	CRI	7.81	96.65
APOG	18.40	97.17	KNOL	12.44	97.73	CBT	7.65	97.31
ISRG	18.22	97.04	BIIB	12.43	94.15	FCN	7.64	97.10
CSE	17.04	93.00	MRTN	12.41	98.43	AYI	7.59	98.24
INTC	16.94	93.92	RIGL	12.20	94.71	CR	7.35	97.61
GE	16.93	94.35	SJW	12.15	99.04	CCO	7.17	96.75
GPS	16.73	95.10	EWBC	12.02	95.11	MXWL	7.12	96.14
DK	16.56	98.08	SF	11.92	97.71	CETV	7.00	96.62
KR	16.50	94.95	CTSH	11.88	95.80	GAS	6.60	97.36

CPSI	16.36	98.30	FPO	11.79	97.75	ESRX	6.51	95.75
ISIL	16.30	93.57	AMED	11.73	96.05	NUS	6.23	96.34
PG	15.90	94.99	CBZ	11.62	95.79	AGN	6.13	97.42
GILD	15.87	92.05	MFB	11.40	98.61	PBH	5.85	94.83
CB	15.81	97.10	AINV	11.12	94.02	CDR	5.47	95.09
MOS	15.77	96.44	BZ	11.05	93.74	NSR	5.07	97.09
ADBE	15.62	94.75	MELI	10.95	97.21	PTP	4.63	97.43
DIS	15.43	94.87	ARCC	10.84	95.17			
EBF	15.26	98.91	ANGO	10.68	96.95			

Table 3: Position of Fleeting Orders

This table presents the position of order placement for orders with a life of 50 milliseconds or less.

Position of Fleeting Orders	Percentage
Inside the bid and ask	11.25
At the best bid and ask	52.23
Less than 10 cents away from the best bid and ask	29.57
10 cents away from bid and ask but not stub quotes	6.93
Stub quotes (buy with a price less than 75% of the bid and sell with a price greater than 125% of ask	0.03

Table 4: Channel Factor Regression

This table presents the summary of the result 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, $j \in \{1,2,3,4,5,6\}$ represents one of the six channel indices of NASDAQ. $f_{i,t}$ stands for the number of the message flow for each stock at time *t*. $marketmessage_t$ is the message flow for all the NASDAQ stocks at time *t*, $residualchannel_{jt}$ is the residual for regressing message flow of Channel *j* on 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 $\gamma_{i,j}$ 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 co-movements with a different channel. The *t*-statistics for the hypothesis that $\gamma_{i,j} = 0$ are in the parentheses. ***, **, * is statistical significance at the 1%, 5%, and 10% levels.

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 1Residual	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 5: Discontinuity Test

This table presents the result on discontinuity test. Panel A lists stocks used for discontinuity test: based on the alphabetical order, they are the first and last stock in each channel with a minimum of one message in each 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 6: Contribution of Fleeting Orders to Liquidity

This table compares the transaction cost and depth of the full limit order book and a limit order book without orders with a life less than or equal to 50 milliseconds. The sample period is from March 19, 2010 to June 7, 2010. There are 118 stocks in the sample and each stock date is an observation.

	Full Book	Without Fleeting	Difference	Diff in Percentage
Quoted Spread in Cents				
Mean	0.0597	0.0600	-0.000251	-0.215%
Median	0.0281	0.0281	-0.0000378	-0.116%
Effective Spread in Cents				
Mean	0.0363	0.0367	-0.000399	-0.879%
Median	0.0185	0.0187	-0.0000950	-0.466%
Depth at Best Ask in Shares				
Mean	2084.746	2080.787	3.959	0.109%
Median	271.342	270.636	0.241	0.070%
Depth at Best Bid in Shares				
Mean	2094.613	2091.022	3.591	0.104%
Median	269.432	269.250	0.219	0.064%
Depth Within 10 Cents of Best Ask				
Mean	23283.810	23255.850	27.962	0.071%
Median	2710.993	2710.621	0.543	0.017%
Depth Within 10 Cents of Best Bid				
Mean	23585.150	23557.090	28.0614	0.082%
Median	2618.659	2618.542	0.560	0.017%

Table 7: Contribution of Fleeting Orders to Price Efficiency

This table compares the one minute volatility and the variance ratio of the full limit order book and a limit order book without orders with a life less than or equal to 50 milliseconds. The sample period is from March 19, 2010 to June 7, 2010. There are 118 stocks in the sample and each stock day is an observation. ***, ** and * represent significance at the 1%, 5%, and 10% levels.

Panel A: Stock Date Comparison

One Minute Volatility			
	Full Limit Order Book>Book without Fleeting Orders	Full Limit Order Book=Book without Fleeting Orders	Full Limit Order Book<Book without Fleeting Orders
Number of Cases	1,657	2,774	2,010
Percentage	25.73%	43.08%	31.21%
Variance Ratio			
	Full Limit Order Book Closer to Random Walk	Same	Book Fleeting Orders Closer to Random Walk
Number of Cases	1,797	2,739	1,903
Percentage	27.93%	42.52%	29.55%

Panel B: Statistical Tests

	Full Limit Order Book	Limit Order Book Without Fleeting Order	Differences	*P*-value
One-Minute Volatility				
Mean with *t*-test	0.00124	0.00125	-0.00000247***	0.000
Median With Signed Rank Test	0.0010046	0.0010057	-0.0000009***	0.000
Variance Ratio (Raw Value)				
Mean with *t*-test	0.951	0.95	0.000380*	0.0818
Median With Signed Rank Test	0.956	0.956	0.000363	0.236
Variance Ratio (Measured as the Deviation from One)				
Mean with *t*-test	0.111	0.111	-0.000180	0.335
Median With Signed Rank Test	0.0848	0.0844	0.000367	0.230

Table 8: Effect of Technology Shocks for Liquidity

The table presents the event study for the technology shocks for the four liquidity measures. For each stock and each day, *qt_spread* is the time-weighted quoted spread, *sz_wt_eff_spread* 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 and for each stock. Standard errors are in parentheses, and ***, ** and * represent significance at the 1%, 5%, and 10% levels.

	(1)	(2)	(3)	(4)
Variables	*qt_spread*	*sz_wt_eff_spread*	depth	depth10
after	-0.000394	0.0000115	-68.31	-2,015***
	(0.00124)	(0.000301)	(93.46)	(736.50)
logvol	-0.00418***	-0.000713**	-114.60	-5,317***
	(0.00147)	(0.000358)	(111.30)	(877.20)
prc	0.000907***	0.000234***	25.42**	118.3
	(0.000141)	(0.0000343)	(10.66)	(83.98)
range	0.0167***	0.00441***	126.90**	-1,057**
	(0.000793)	(0.000193)	(59.91)	(472.10)
Constant	0.0596***	0.0127**	5,001***	118,697***
	(0.0211)	(0.00512)	(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 9: Effect of Technology Shocks on Price Efficiency and Volume

The table presents the event study for the technology shocks on price efficiency and volume. For each stock and each day, volatility is the one-minute volatility, variance is the one-minute variance ratio, and volume is the daily volume.

Variables	(1) *sigma_all*	(2) *all_ratio*	(3) volume
after	0.0000249 *	-0.00289	131,609
	(0.0000128)	(0.00332)	(142,487)
Constant	0.00114***	0.951***	5.971e+06***
	(9.04e-06)	(0.00234)	(100,625)
Observations	5,858	5,856	5,860
R-squared	0.001	0.000	0.000
Number of ticker	118	118	118

Standard errors are in parentheses.
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Figure 1: Histogram of Quote Life for Orders with a Life Less than One Second

This graph presents the histogram for all the orders with a life less than one second. Each bin represents a 5-millisecond interval. The sample includes 118 stocks from March 19, 2010 to June 7, 2010.

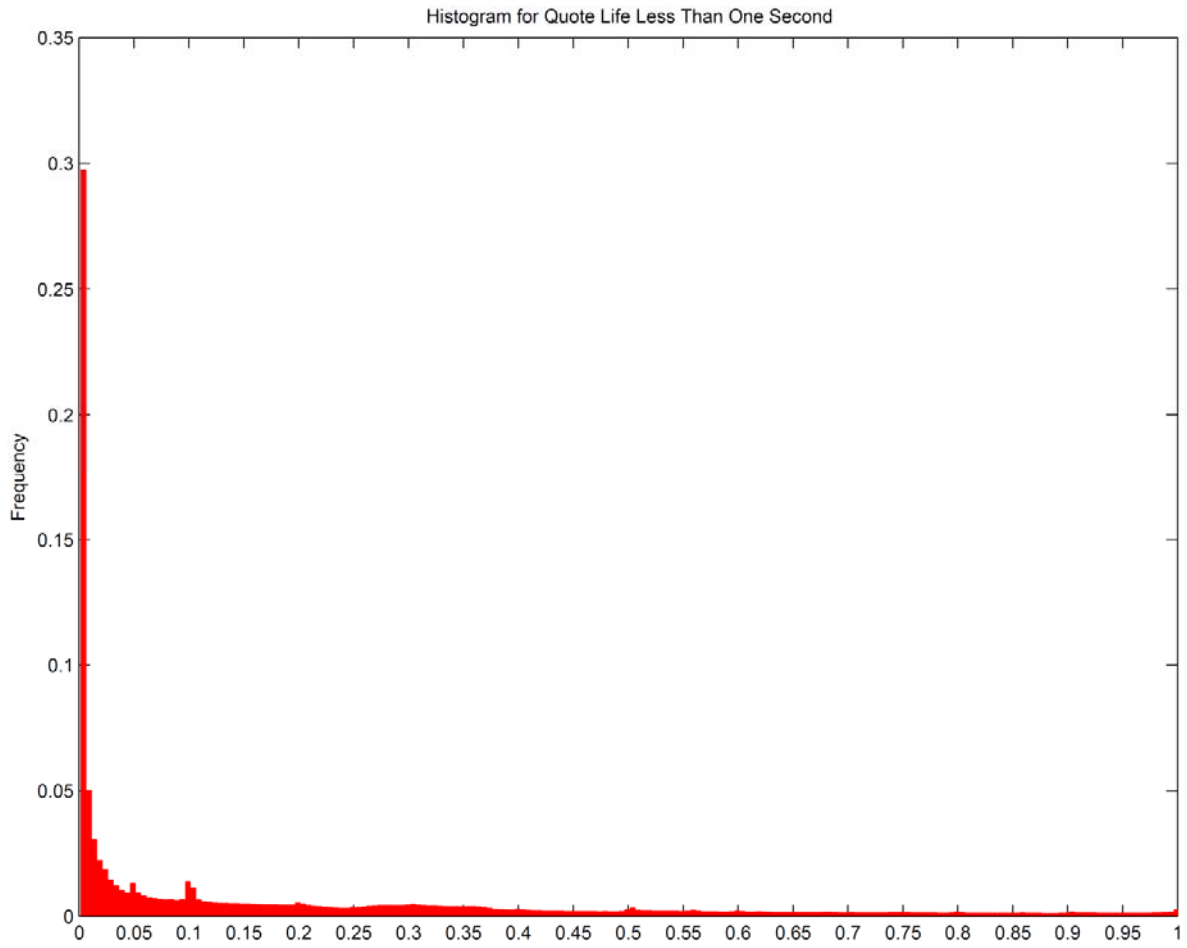

Figure 2: The Impact of Technology Shocks on Latency

This figure demonstrate the impact of are two technology shocks on latency. The first technolgy shock happened between Friday, April 9, 2010 and Monday, April 12, 2010. The second happened between May 21, 2010 and May 24, 2010. We have two measures of latency. Panel A domonstrates the minimum time differences between two consecutive messages for the NASDAQ market. Panel B domonstrates the fastest cancellation and execution for the NASDAQ market.