The Externalities of High-Frequency Trading

March 15, 2012

Abstract

We show that two exogenous technology shocks that increase the speed of trading from microseconds to nanoseconds do not lead to improvements on quoted spread, effective spread, trading volume or variance ratio. However, cancellation/execution ratio increases dramatically from 26:1 to 32:1, short term volatility increases and market depth decreases. We find evidence consistent with "quote stuffing," which involves submitting an extraordinarily large number of orders followed by immediate cancellation in order to generate order congestion. The stock data are handled by six independent channels in the NASDAQ based on alphabetic order of ticker symbols. We detect abnormally high levels of co-movement of message flows for stocks in the same channel using factor regression, a discontinuity test and diff-in-diff test. Our results suggest that an arms race in speed at the sub-millisecond level is a positional game in which a trader's pay-off depends on her speed relative to other traders. This game leads to positional externality (Frank and Bernanke, 2012), in which private benefit leads to offsetting investments on speed, or effort to slow down other traders or the exchange, with no observed social benefit.

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]

The empirical literature on the speed of trading before the sub-millisecond era finds the social value of increases in speed. For example, Hendershott, Jones and Menkveld (2011) find that the automated quote dissemination in the NYSE reduces the spread and enhances the informativeness of quotes in 2003. In contrast to the previous work, this paper shows that such a benefit has ceased when the speed improvement proceeds to the micro or nano second level. Two exogenous technology shocks that increase the speed of trading from microseconds to

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

nanoseconds do not lead to improvements on market quality 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 lead to a dramatic increase in the cancellation/execution ratio from 26:1 to 32:1 and an increase in short term volatility as well as a decrease of market depth.

As the speed provides private value to a trader, it is equally valuable to slow down her competitors. Biais and Woolley (2011) discuss a trading strategy called "[quote] stuffing," a type of externality-generating behavior, which involves submitting a profuse number of orders to the market to generate congestions on purpose. Though regulators classify quote stuffing as a type of market manipulation,[3] the behavior itself is hard to identify. For example, Egginton, Van Ness, and Van Ness (2011) find that intense quoting activity is correlated with short-term, but it lacks convincing evidence of their causal relationship. It is even less clear to identify whether the intense episodic spikes of quoting activity are generated through manipulative "quote stuffing" or they are natural responses to a market with higher short-term volatility.

This paper provides a clear identification strategy for quote stuffing activities based on channel assignments of NASDAQ-listed stocks. The data feed for NASDAQ-listed stocks are divided into six identical but independent channels.[4] Trading data is split between the six

[3] 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."

[4] 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.

channels based on the first character of the issue symbol.[5] The channel assignment is close to random with respect to firm fundamentals, thus providing us with a clean identification scheme for one type of quote stuffing.[6] [7] Excessive message flow of a stock stifles the trading of stocks in the same channel, but it does not have the same effect on stocks in a different channel. 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 test the quote stuffing behavior based three methodologies: first we show the existence of abnormal message flow co-movement for stocks handled by the same channel through 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 [Lessard (1974, 1976), Roll (1992), Heston and Rouwenhorst (1994), Griffin and Karolyi (1998), Cavaglia, Brightman, and Aked (2000), and Bekaert, Hodrick, and Zhang (2009)]. In our application, the six channels in total resemble a "global market," whereas each

[5] 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] This type of quote stuffing affects the consolidated data feed. Most traders in the market use a consolidated data feed. Some high-frequency traders may subscribe to the direct data feeds for some market centers while using a consolidated data feed for 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. 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.

[7] 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 time to figure out which gateway is open.

channel represents a “country”. The factor regression reveals a diagonal effect: after controlling for the message flow of the “global” market, the message flow of a stock has an abnormal positive correlation with the total message flow of other stocks in its own channel. In contrast, the message flow of a stock has negative correlations with message flows of stocks in all other channels. Our second identification method, a discontinuity test, also demonstrates the positive abnormal correlations of message flows of stocks handled by the same channel. 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.[8] Our third identification method, a diff-in-diff regression, further strengthens the results. 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.

Our result elicits an intuitive economic interpretation. The level of bid-ask spread is related to the liquidity providing function of high frequency trading. Current U.S. stock markets observe price, display and time priority.[9] The fierce competition in speed implies the failed competition in price. The fact that an increase in speed does not change the bid-ask spread supports this hypothesis. In other words, high frequency traders cannot undercut each other by

[8] 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.

[9] 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.

price, but the faster trader can eventually provide liquidity because of his earlier arrival than other traders. In the standard definition of 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 the market at the same time. The reality in the financial market, however, is exactly the opposite, where time becomes divisible at the nanosecond level but price is restricted by tick size. Therefore, suppose that zero profit (or equilibrium) bid-ask spread is 1.5 cents. Then, the liquidity provider will lose money if he chooses a bid-ask spread of 1 cent, but there exists abnormal profit if he sets the bid-ask spread to be 2 cents. The 0.5 cent rent per share provides incentive for competing in speed. This competition has two implications.

First, competition in speed but not price matches the definition of externality (Laffont, 2008).[10] By increasing speed, a high frequency trader directly harms the production set of liquidity of his competitors. The private benefit of increasing speed for one high frequency trader is higher than the social benefit, because part of profit earned by the faster trader is "stolen" from slower high frequency traders. Aghion and Howitt (1992) term this externality "business stealing effect." A more general discussion of the consequence of this externality can be found in the canonical textbook by Tirole (1988).[11] Basically, a firm that invests in speed does not internalize the loss of profit suffered by its rivals, which suggests too much investment on speed in equilibrium. Most important of all, competition in speed does not work through the price system. In fact, it is the failure of price competition that leads to speed competition. Competition working through price system does not lead to externality, because the loss to producers is precisely offset by the gain to consumers (Laffont, 2008). Competition in speed, however, does not have such

[10] 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. (New Palgrave Dictionary of Economics, second edition)

[11] Answer for exercise 10.5 in page 416 of the book demonstrates mathematically the magnitude of the externality and also offers the economic intuition.

effect unless the consumer of liquidity cares directly about the difference between micro and nanoseconds.

Secondly, quote stuffing provides strong evidence that competition in speed is a positional game, in which a trader's pay-off depends on his speed relative to other traders. 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 to a greater extent. Recent work by Frank (2003, 2005, 2008) and Bernanke and Frank (2010) argue that positional games lead to positional externality, because any step that improves one side's relative position necessarily worsens the other's ranking. In our case, quote stuffing creates benefit to the initiator, but there is no social benefit associated such activity.

More importantly, speed competition imposes negative externalities to traders who are not in the speed game. An increase in speed decreases the quoted depth and increases short term volatility of price. In addition, order cancellation increases despite of steady trading volume, which implies that the size of the data increases. We believe that the increase in speed leads to more discrete time periods for a fixed calendar time, which increases the number of possible moves for a trading game among high frequency traders. The game among high-frequency traders becomes more complex, but the aggregated opportunity for actual trading with non-high frequency traders is unlikely to increase. As a result, we witness an increase in cancellation and short-term volatility. Depth also decreases, probably because it becomes more risky to expose a large size order when increases in speed increase pick-off risk. We show that order cancellations now consume 97% of computer system resources, which the entire market has to bear.[12] The high levels of cancellations force stock exchanges and traders to continually upgrade trading

[12] According to Wharton Research Data Services, the Trade and Quote Data (TAQ) is more than 10 terabytes per year, the same size as the digitized versions of all prints in the Library of Congress.

systems and bandwidth to accommodate higher message flows. In addition, most stock exchanges only charge fees for executions but not cancellations. This worsens the externality problem because traders who actually execute orders are subsidizing those traders with excessive cancellations.

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), 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, which increases social welfare; on the other side, high-frequency trading may cause an adverse selection problem and generate negative externalities to traditional traders and investors. While an increase in speed from seconds to milliseconds may result in more trading opportunities, our results 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. The identification problem naturally arises due to the endogenous relationship between liquidity, price discovery, order cancellations, and speed. Computing power also presents a serious challenge.[13] We address the identification issue based on two exogenous technology shocks and NASDAQ channel assignments. These two identification strategies are implemented by two supercomputers from the National Science Foundation's Extreme Science and Engineering

[13] 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.

Discovery Environment (XSEDE) program. To our knowledge, our empirical investigation is one of largest computing efforts ever conducted in academic finance.

More broadly, our paper is related to the literature of overinvestment in research and development, information acquisition, professional services, and financial expertise. 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 the discovery of hidden properties of nature that can only be laid bare by action. We conjecture that the information existing at the microsecond or nanosecond level is more of the former. 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. The general notion that agents may overinvest to compete in a zero-sum game links back to Ashenfelter and Bloom (1993). A more recent work by Glode, Green, and Lowery (2011) examines the arms race for financial expertise.

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 the NASDAQ. In Section 5, we use event studies to compare the market quality before and after the system enhancements of speed. Section 6 concludes the paper and discusses possible policy implications.

2. Data

2.1 NASDAQ TotalView-ITCH Data

The main dataset for this paper is the NASDAQ TotalView-ITCH, which is a series of messages that describe orders added to, removed from, and executed on the NASDAQ. The data come as a daily binary file and the 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 a 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 for all the NASDAQ-listed 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 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 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. A U message 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 a U message. In short, new orders can originate from three message files: messages A, F, and U.

A 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 a 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.

A 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 uses to have 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, a D message 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.

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 2,377 common stocks listed on the 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). Since our sample period extends to 2011 and Hendershott and Riordan picked the stocks in early 2010, 118 of the 120 stocks remain in the sample.

With the help of the 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 usually when both the exchanges and traders test new technology. The first structural break happened between Friday, April 9, 2010 and Monday, April 12, 2010. A more dramatic change happened between Friday, May 21, 2010 and Monday, May 24, 2010. 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 the fastest trader is so large that it is beneficial to implement and use the innovation once it is mature. Figure 1 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 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 500-600 nanoseconds and stays below one microsecond for all but seven days after the break. Therefore, NASDAQ enters the realm of nanosecond trading after May 24, 2010.

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 the administrative information when we construct the limit order book, but we do not count the stock symbol and administrative information in the total message flow. The result is similar even if they are added 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:

$$\text{Cancellation\_ratio} = 1 - \frac{E+C}{A+F+U}. \tag{1}$$

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

$$\textit{Cancellation\_execution} = \frac{D+X+U}{E+C}. \tag{2}$$

The U type message is in both definitions because a U message involves a deletion plus an addition. 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 messages and order deletion through D, X or U messages. 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 than the number of cancelled or updated orders.

We also use A, F, U, E, C, X, and D messages 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 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 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. 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 midpoint 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. [14]

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.

3 Preliminary Results

Table 2 presents the order cancellation ratio. NACCO Industries (Ticker NC) has the highest cancellation ratio, with 99.57% of submitted orders cancelled. Some of the most liquid stocks have very high cancellation ratios. For example, 96.09% orders of Apple (AAPL) are cancelled and 95.92% of Google (GOOG) orders are cancelled. 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.

Figure 2 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

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

stocks for which we construct the limit order books. 30% of the observations fall into the bin with the shortest quote life. This result has the following implication. Regulators across the Atlantic are proposing minimum quote life policy to slow down the trading process. In Europe, the *Review of the Markets in Financial Instruments* (MiFID) solicits comments on "How should the minimum period be prescribed?"[15] In the United States, "The likely minimum duration for a quote under such a proposal could be 50 milliseconds, which has been suggested by several sources."[16] Currently, the minimum quote life for most actively traded foreign exchange currency pair is 250 milliseconds.[17] Our paper does not directly address the minimum quote life policy, but we define order with a quote life less than 50 milliseconds as fleeting orders. Figure 2 demonstrates that a minimum quote life of 50 or 250 milliseconds would not generate a significant difference in market outcome because there are few observations in-between.

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 fleeting orders are placed inside BBO, while 52.23% are placed at the BBO and 36.53% are placed outside the BBO,[18] which suggests that fleeting orders are placed for different purposes in 2010 than in 2004.

We evaluate the contribution of fleeting orders to liquidity and price discovery by constructing two limit order books: one with all orders and one excluding orders with a life less than 50 milliseconds. We call this partial equilibrium analysis because we do not consider the

[15] *European Commission Public Consultation: Review of the Markets in Financial Instruments Directive* (*MiFID*), February, 2011, page 7.

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

[17] Thomson Reuters Spot Matching: Changes to Minimum Quote Life and Transaction to Match Ratio, October 17, 2012.

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

complex dynamics if the SEC enforces a 50-millisecond minimum quote life. We supplement the partial equilibrium analysis with a natural experiment in the Section 5.

The results for our four liquidity measures of the 118 stocks for 55 days between March 19, 2010 and June 7, 2010 are shown in Table 4. The daily measure for one stock is an observation. Table 4 shows that the average quoted spread for the whole book is about 5.97 cents and the median is about 2.81 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 6.00 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 0.215% to 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 0.00378 cents in terms of median spread, which is 0.116% of the liquidity. The result for the effective spread is slightly larger: the limit order book without fleeting orders has a 0.0399 cent increase in the effective spread in terms of mean spread and a 0.0095 cent increase in terms of 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 28 shares to the depth within 10 cents of the best bid and ask, and the median number is 0.54 shares for the 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.

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 5 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. The variance ratio results are neither economically nor statistically significant. In fact, if we measure the difference between the variance ratio and 1, the test based on mean suggests that the return in the full limit order book is closer to a random walk, while the test on the median suggests 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.

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. 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 microsecond or nanosecond trading environment, it is 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 exchanges, or both. The economic incentives for enhancing speed and delaying others should be the same, if it is relative speed that is important. According to Brogaard (2011c), 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 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 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 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).[19] 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 explained by footnotes 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 UTP data because the channel assignment provides us with the identification strategy.

[19] 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 (CTS).

Suppose, for example, a trader has information for Stock A. One way 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 likihood 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 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.

4.1 Factor Regression

We obtain the channel assignments for NASDAQ-listed stocks from the 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 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 six channels as the "global market." The literature on country factor examines whether there is a country specific factor after controlling for the global market co-movement. Using the same method, 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. We make this adjustment to avoid mechanical upward bias to find that a stock has higher correlations with message flows in its own channel. The market message flow is the sum of the messages for all stocks.[20] For each stock i, we run the following two stage regressions following Bekaert, Hodrick, and Zhang (2009)[21]:

We first regress the total number of messages of Channel j on the market message flow:

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

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:

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

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 6

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

[21] 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.

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 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 coefficients are zero. 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 with the message flow for the channel even after controlling for the 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 a 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 use the first stock.[22] 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:

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

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

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 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 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.

4.3 Diff-in-diff Regression

Our final test for abnormal co-movement for message flow is based on 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 control group changes ticker symbols but remains in the same channel; the treatment group changes 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 switch 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 of ticker symbol.   For each stock, we use the 30 days before the ticker change as before period and 30 days after the ticker change as after period.

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 1 percent 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.

5. Natural Experiment

To evaluate the effects of the technology shocks on liquidity, price efficiency and trading volume, 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. We compare the market liquidity and price efficiency before and after these two technology shocks. These two structural breaks, particularly the one happened in May 21, 2010, dramatically increases the trading speed. It also increases the cancellation ratio. For the event days before and after these structural changes, the mean cancellation/execution ratio increases from 25.82 to 32.04, while the cancellation/execution ratio increases from 20.30 to 33.56 between March 2010 and June 2010.

5.1 Effects of the Technology Shocks on Liquidity

Following the approach of Boehmer, Saar, and Yu (2005) and Hendershott, Jones, and Menkveld (2011), we regress the liquidity measure L_{it} on the event dummy and a number of controls. Our liquidity measure includes (time weighted) quoted spread, (size weighted) effective spread, and (time weighted) depth that at the best bid and ask and (time weighted) depths within 10 cents of the best bid and ask.

$$L_{it} = \mu_i + \alpha After_t + \beta_1 logvol_{it} + \beta_2 range_{it} + \beta_3 Prc_{it} + \varepsilon_{it,} \tag{6}$$

$log\ vol_{it}$ 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 day high minus 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 9 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 of the best bid and ask. Overall, we find that these two technology shocks neither increases nor decreases spread but slightly decrease the depth.

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 very little room for competition for price at the best bid and asks. As a result, spread can barely decrease when speed increases. Second, one argument that speed may increase liquidity is that traders with high speed can maintain 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 are unlikely to change. We also find that market depth slightly decrease, probably because it is more risky to expose a large position when speed is higher.

5.2 Effects of Technology Shocks on Market Efficiency and Volume

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.

$$E_{it} = \mu_i + \lambda After_t + \varepsilon_{it,} \tag{7}$$

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.

Table 10 shows that the variance ratio at 1 minute level does not have a statistically significant change before and after the technology shocks. The change of trading volume is also not statistically significant. However, volatility slightly increases after the technology shocks.

5.3 Summary

We find that two exogenous technology shocks do not affecting volume, spread and variance ratio. However, it dramatically increases cancellation/execution ratio and increases short term volatility and decreases market depth. We believe that an increase in trading speed increases the number of periods for the trading game played between high-frequency traders. Therefore, we see more order cancellations, probably because a more complex game leads to higher cancellations. For example, the quote stuffing strategy may need increasingly more orders to generate congestion. However, an increase in speed does not improve liquidity or price efficiency.

As a result, speed may create several externalities. Quote stuffing is certainly one type of externality-generating events. Even without quote stuffing, we argue that investment in speed with sub-millisecond accuracy may provide a private benefit to traders without consummate social benefit; therefore, there may be an overinvestment in speed. Finally, the exchanges continually makes costly system enhancements to accommodate higher message flow, but these enhancements facilitate further order cancellations, not increases in trading volume. Since the current exchange fee structure only charges executed trades and not order cancellations, legitimate traders and investors subsidize high-frequency traders who purposefully cancel orders, reflecting a wealth transfer from low frequency traders to high-frequency traders.

6. Conclusion

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

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.

We also find that fleeting orders, or orders with a life less than 50 milliseconds, have a trivial contribution to liquidity and no contribution to price efficiency. We also 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. Spread does not decrease following 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 finding is consistent with the observations from the market on the accumulative effects of a series of enhancement in speed. U.S. Securities and Exchanges Commission (2010) reveals that the average trade size has decrease 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.

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. 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, this would be 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.

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 nanosecond, 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 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. 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.[23] 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."[24] 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

[23] 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.

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

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.

References

Aghion, Philippe and Peter Howitt 1990. A model of growth through creative destruction NBER Working Paper

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

Frank, Robert. 2003. Are positional externalities different from other externalities? Working Paper.

Frank, Robert. 2005. Positional externalities cause large and preventable welfare losses. The American Economic Review. 95(2) 137-141.

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

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

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.

Grossman, Gene M. and Elhanan Helpman. 1991. Quality ladders in the theory of growth. The Review of Economics Studies. 58(1), 43-61.

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, 2011, 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, 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.

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	Timestamp	Order					Original Order	Match	Market
Type	(nanoseconds)	Reference Number	Buy/Sell	Shares	Stock	Price	Reference Number	Number	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: Percentage of Fleeting orders and the level of Cancellation

This table presents the percentage of orders cancelled (Cancel Ratio).

Stock	Cancel Ratio	Stock	Cancel Ratio	Stock	Cancel Ratio	Stock	Cancel Ratio
NC	99.57	PTP	97.43	MAKO	96.42	PBH	94.83
ERIE	99.56	AGN	97.42	CNQR	96.35	HPQ	94.8
CRVL	99.49	ROC	97.4	NUS	96.34	ADBE	94.75
ROG	99.42	DCOM	97.39	KMB	96.32	CPWR	94.73
AZZ	99.29	ROCK	97.37	LPNT	96.3	RIGL	94.71
PPD	99.22	GAS	97.36	MMM	96.25	FULT	94.66
SJW	99.04	CBT	97.31	MXWL	96.14	CMCSA	94.59
EBF	98.91	MANT	97.29	AAPL	96.1	FMER	94.46
CKH	98.75	JKHY	97.22	AMED	96.05	GE	94.35
BW	98.69	MELI	97.21	FRED	95.95	GLW	94.18
MFB	98.61	APOG	97.17	GOOG	95.92	BIIB	94.15
IPAR	98.49	CB	97.1	CTSH	95.8	IMGN	94.07
MRTN	98.43	FCN	97.1	ABD	95.79	AINV	94.02
LECO	98.37	NSR	97.09	CBZ	95.79	INTC	93.92
SFG	98.32	ISRG	97.04	ESRX	95.75	CSCO	93.91
LANC	98.3	ANGO	96.95	CBEY	95.59	PFE	93.86
CPSI	98.3	RVI	96.87	AXP	95.57	BRCM	93.77
AYI	98.24	KTII	96.82	BAS	95.55	BZ	93.74
DK	98.08	CCO	96.75	COST	95.54	GENZ	93.7
FFIC	98.02	MOD	96.74	FL	95.31	DELL	93.68
CTRN	97.8	BRE	96.72	ARCC	95.17	CELG	93.68
FPO	97.75	AMZN	96.68	EWBC	95.11	ISIL	93.57
KNOL	97.73	CRI	96.65	SWN	95.1	AMAT	93.22
SF	97.71	CETV	96.62	GPS	95.1	EBAY	93.1
CSL	97.67	HON	96.61	CDR	95.09	CSE	93
CR	97.61	LSTR	96.61	DOW	95.01	AMGN	92.95
PNY	97.6	NXTM	96.56	PG	94.99	MDCO	92.92
COO	97.58	MIG	96.47	KR	94.95	GILD	92.05
BXS	97.47	BHI	96.44	AA	94.92		
PNC	97.43	MOS	96.44	DIS	94.87		

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: Contribution of Fleeting Orders to Liquidity

This table compares the transaction cost and depth of the full limit order book and the 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	5.97	6.00	-0.0251	-0.215%
Median	2.81	2.81	-0.00378	-0.116%
Effective Spread in Cents				
Mean	3.63	3.67	-0.0399	-0.879%
Median	1.85	1.87	-0.00950	-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 5: 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 the limit order book without orders having a life less than 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. respectively.

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 (Measured as the Deviation from 1)				
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 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 kth column and the jth 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 that 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 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 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 control group changes ticker symbol but remain in the same channel; the treatment group changes 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)

Table 9: Effect of Technology Shocks for Liquidity

The table presents the event study of the technology shocks for the four liquidity measures. For each stock per 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 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	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 10: Effect of Technology Shocks on Price 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, variance is the one-minute variance ratio, and volume is the daily volume.

	(1)	(2)	(3)
Variables	sigma_all	all_ratio	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: The Impact of Technology Shocks on Latency

These figures demonstrate the impact of our two technology shocks on latency. The first technology shock happened between Friday, April 9, 2010 and Monday, 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.

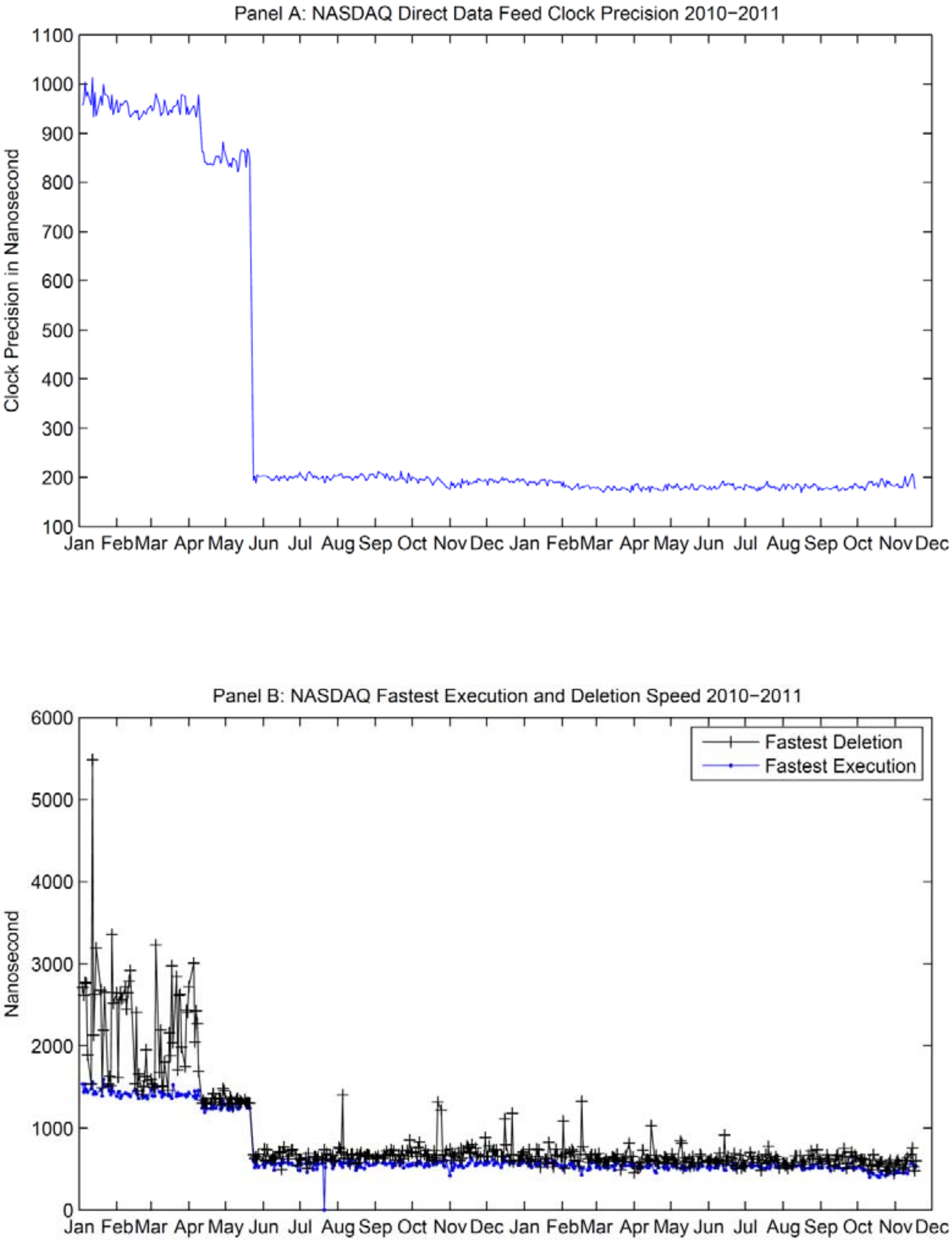
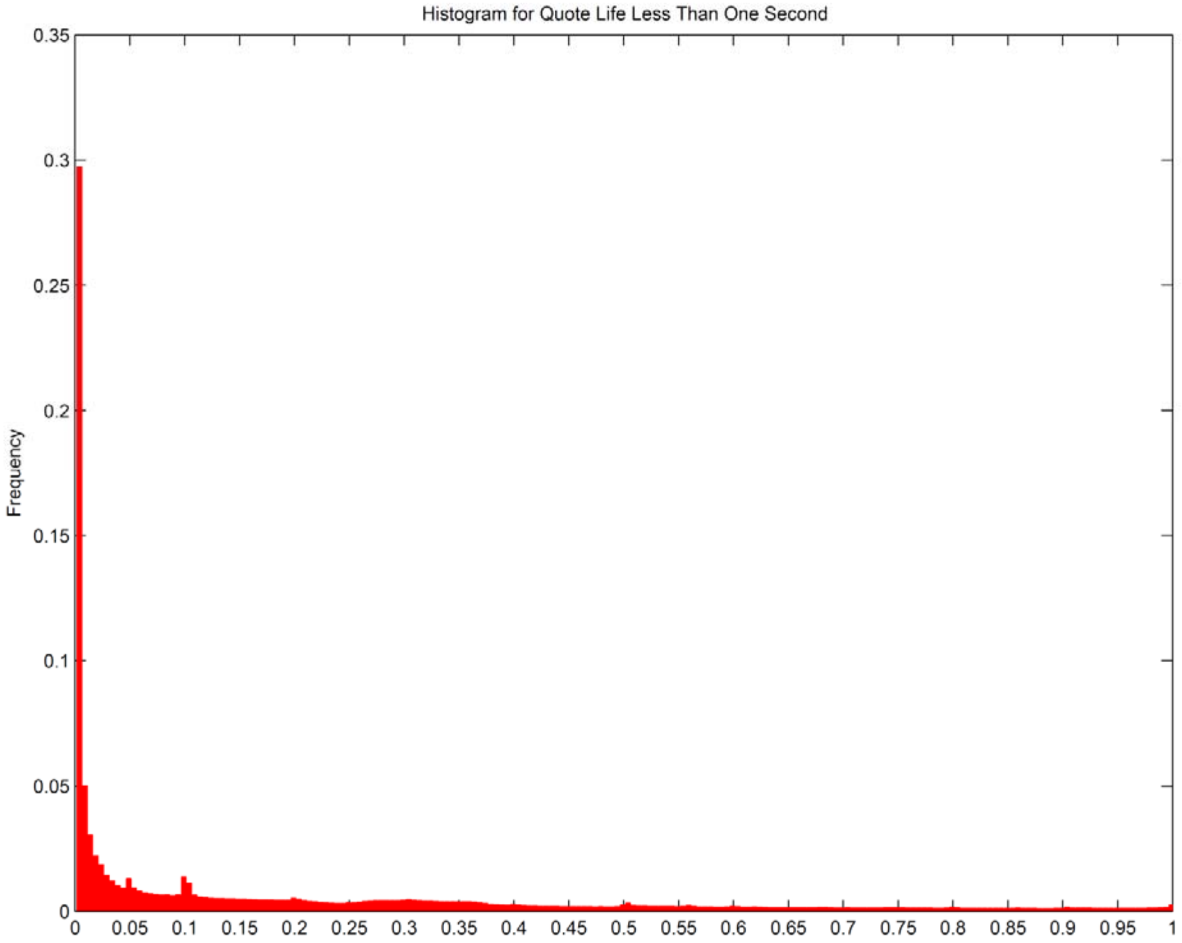


Figure 2: Histogram of Quote Life for Orders with a Life Less than One Second

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


Histogram for Quote Life Less Than One Second
Frequency