Price Constraint, Competition in Speed and Market Quality[1]

September 30, 2013

Chen Yao Mao Ye

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

Abstract

When price competition is constrained by tick size, speed allocates the resources due to the time priority rule. We find more high frequency liquidity provision for lower price stocks with higher market cap, where the one cent tick size has a higher constraint on the price competition. The same stocks also have higher market shares in markets with inverts fee, where liquidity providers need to pay a fee instead of getting a rebate. We find that exogenous technology shocks improving speed at millisecond, microsecond or nanosecond level mostly affect the trading speed of stocks that have already had constrained price. As a consequence, speed competition does not change quoted spread, effective spread and variance ratio. However, short term volatility increases and market depth decreases. Instead of discussing whether and how we should regulate high frequency trading and maker/taker fee, our results suggest these two phenomena may be consequences of regulation itself (SEC rule 612, or Minimum Pricing Increment Rule). One possible policy solution is deregulation.

Key Words: Price Constraint, High-frequency trading, marker/taker fee, liquidity, price efficiency, tick size

1. Introduction

In standard Walrasian equilibrium, price is infinitely divisible but time is not; all agents are assumed to arrive at the market at the same time. However, the reality of the financial market, is exactly the opposite: time becomes divisible at the nanosecond level but price is restricted by tick size. This paper shows two frictions, discrete price and (almost) continuous time as regards to the canonical Walrasian equilibrium helps explain two most important features of current market: high frequency trading and taker/maker fee.

Recently, *The Wall Street Journal* stated that trading entered the nanosecond age when Fixnetix, a London-based trading technology company, announced "it has the world's fastest trading application, a microchip that prepares a trade in 740 billionths of a second, or nanoseconds." Since "investment banks and proprietary trading firms spend millions to shave ever smaller slivers of time off their activities, ...the race for the lowest 'latency' [continues], some market participants are even talking about picoseconds — trillionths of a second."[2] Because high-frequency traders aggressively invest in technologies that reduce latency, speed should create private benefits. A recent article in Financial Times, estimates that a one millisecond advantage is worth up to $100 million in annual gains.[3] A more important question, however, is whether investment in speed creates social benefit by improving liquidity. The empirical work on the speed of trading before the sub-millisecond era find that speed decreases

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

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

quoted spread and effective spread.[4] Put differently, the enhanced speed of providing or cancelling liquidity leads to better (prices of) liquidity.

However, speed competition, by its nature, is a form of non-price competition. U.S. stock markets observe price, display and time priority.[5] If liquidity providers can still undercut each other by price, speed competition is secondary. Therefore, fierce speed competition may not facilitate price competition. On the contrary, we argue that speed competition in liquidity provision may simply be a consequence of failed price competition. In United States, tick size is regulated through SEC rule 612 (Minimum Pricing Increment) of regulation NMS. The rule prohibits stock exchanges from displaying, ranking, or accepting quotations, orders, or indications of interest in any NMS stock priced in an increment smaller than $0.01 if the quotation, order, or indication of interest is priced equal to or greater than $1.00 per share. Because the tick size is the same for stocks above $1.00, the natural bid-ask spread for the largest stocks with the lowest price is more likely to be constrained by tick size, and we find that these stocks have the highest level of high frequency liquidity provision. By the doubling sorting stocks by market cap and price, we find that large stocks with lower price have 45.4 percent of liquidity provided by high frequency traders, whereas the number for high price stock is only 30.9 percent.

We also identify another measure of price constraint in addition to price. Although stock exchanges cannot quote sub-penny, this rule can be by-passed by the taker/maker fee. For example, EDGA and EDGX are two trading platforms offered by Direct Edge. They are almost identical except in one dimension. EDGX, like most other exchanges uses maker/taker fee

[4] See Boehmer, Fong and Wu (2013), Hendershott, Jones, and Menkveld (2011), Hendershott and Riordan (2009 and 2011) and Hasbrouck and Saar (2013), among others.

[5] 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.

structure, which pays liquidity makers and charges liquidity takers. EDGA, however, has an inverted taker/maker fee structure that charges liquidity providers and pays liquidity takers. The fact that liquidity providers are willing to pay a fee to make the market implies that nominal bid-ask spread is constrained. The more the price constraint, the more active the market with a taker/maker fee. Therefore, the market share of EDGA to EDGX is a natural measure of price constraint. This proxy is further justified by the theoretical model of Foucault, Kadan and Kandel (2013). In their model, stock exchanges find it optimal to charge liquidity maker and subsidize liquidity taker when the tick size is large and vice versa. In reality, stocks with different prices have different relative tick sizes, but stock exchanges cannot adjust the fee based on the stock characteristics. The second best is to offer two trading platforms, one with a maker/taker fee and the other with an inverted fee. Stocks with higher price constraint have an active taker/maker fee market whereas stocks with lower price constraint have an active maker/taker fee market. Therefore, the price constraint not only provides explanations for the coexistence of the two markets with opposite fee structures, but also explains their relative market share. We use the market share of EDGA to EDGX as our second proxy for price constraint, and find that it is closely related to our first measure: price and market cap. For each market cap group, EDGA has higher market share to EDGX for stocks with lower prices.

The market share of EDGA to EDGX reveals close association between price constraint and high frequency market making. By sorting the stocks into five portfolios, we find that highest EDGA to EDGX market share quintile has twice as much high frequency market marking activity as the lowest EDGA to EGDX market share quintile. The ratio alone can explain 32% of variation in high frequency market marking.

To our knowledge, our paper is the first empirical work that provides a coherent framework linking the literature of high frequency trading[6] to the maker/taker fee market[7]. Two common factors, price constraint and time priority contributes to these two market phenomena. Higher price constraint implies higher profit of supplying liquidity and an oversupply of liquidity at the constrained price. Traders with higher speed are able to supply liquidity because of the time priority in the queue. Meanwhile, higher price constraint implies that some traders are willing to pay a fee in order to make the market, especially when the market where makers obtain rebates has a long queue. This explains the proliferation of the markets with an inverted fee.

Our paper contributes to the literature by providing a new interpretation to the competition in speed, which offers new insight to the impact of speed competition on market quality. When the speed competition is a consequence of the price constraint, speed may determine who provides liquidity, but it will not change the price of liquidity, because different liquidity providers can no longer undercut each other by price. This result is verified by several major technology shocks that increase trading speed. First, we identify an enhancement of the matching engine of the NASDAQ in April 2010 and a follow-on technology enhancement from the high frequency trader side in May 2010. These two shocks have reduced the latency of submitting and processing orders from micro to nanoseconds but they did not improve liquidity measures. Quoted spread, effective spread, trading volume and variance ratio stay around the same level after the shocks. However, these shocks lead to increased short-term volatility as well as a decreased market depth. We also document technology enhancements that reduce the

[6]See Chaboud, Chiquoine, Hjalmarsson, and Vega (2009), Hendershott and Riordan (2009 and 2011), Brogaard(2011 a and b), Hasbrouck and Saar (2013) and Hendershott, Jones, and Menkveld (201) among others.
[7] See Foucault, Kadan and Kandel (2013), Colliard and Foucault (2012), Brolley and Malinova (2012) Park and Malinova (2013) Halmrast, Malinova and Park (2013), among others.

latency of disseminating trading data from 3 milliseconds to 1 millisecond. Interestingly, NASDAQ trading data are disseminated by six identical but independent channels based on the alphabetic order of ticker symbols. Channel 1, with all the NASDAQ stocks ticker symbols beginning with A and B, was upgraded on October 10, 2011.[8] Channels 2 through Channel 6 were upgraded on October 17, 2011. This staggered technology enhancement provides a clean test to examine the casual relation between speed and liquidity. Again, we find that the enhancement of speed does not improve liquidity.

Our result on market quality contrast to the existing work between high frequency trading and market quality, which finds that high frequency trading improves liquidity. However, current empirical evidence is based on data before sub-millisecond era. The speed competition in nanoseconds not only differs in magnitude but also in nature. The current discussion, either in academic literature or in the public domain, focus on the speed advantage between high frequency traders and slow traders such as humans. Speed competition at nanosecond level, however, can only be possible among different high frequency traders. One question remained unanswered is the nature of the mechanism by which the interaction of high frequency trading algorithms improves market quality (Chordia, Goyal, Lehmann and Saar, 2013). Price constraint points the nature of speed competition and its impact on market quality. We find that during the technology enhancement, the stocks with the largest increase in speed are the stocks with price constraint. However, for stocks with price constraint, bid-ask spread cannot be further improved. Therefore, this result not only verifies our hypothesis that speed competition is a consequence of

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

failed price competition, but also implies that speed competition cannot lead to improvement in liquidity.

Our paper also provides new insight to regulations. Instead of discussing whether and how we should regulate high frequency trading and maker/taker fee, our paper is the first one to state these two phenomena may be a consequence of existing regulations. Economic reasoning and our empirical evidences can show, step by step, how different regulations create the world we observe. With infinitely small tick size, the breakdown between maker and taker fee does not matter, because these fees will be neutralized by the differences in the nominal bid-ask spread: The maker/taker market will have a lower nominal spread and the taker/maker market will have a higher nominal spread, but the spread cum fee bid-ask spread are the same for both markets (Angel, Harris and Spatt (2011) and Colliard and Foucault (2012). However, SEC rule 611 states that orders should be routed to the market with the best displayed (nominal) spread. Then, all orders should be routed to the maker/taker market and the taker/marker market should be empty. However, because of the prohibition of sub-penny pricing by rule 612, maker/taker and taker/maker market can display the same nominal bid-ask spread.[9] In addition, rule 611 only imposes price priority across markets but time priority is only imposed in the individual market [10]. With price constraint, the queue in maker/taker market can be very long and the order execution becomes a privilege to the liquidity providers with higher speed. Therefore, some liquidity providers are willing to pay a fee to jump ahead of the queue by trading in taker/maker fee market.

[9] For example, if the equilibrium spread without tick size is 0.3 cents for maker/taker market and 0.8 cent for taker/maker market, both markets will quote 1 cent spread due to price constraint.
[10] If global time priority had been imposed, liquidity providers would abandon the taker/maker fee market to collect rebate in maker/taker market because the execution probability is the same.

Therefore, one possible policy solution for the high frequency trading, maker/taker and taker/maker market is deregulating the market. Reducing tick size will lead to more competition in price instead of the arms race in speed. Certainly, the benefit of reducing tick size needs to be weighed against the adjustment cost, but we are currently observing a reverse trend. SEC is now considering increasing the tick size[11]. The rationale is that "Market makers and dealers need more economic incentives to bring smaller companies public, provide bids and offers and publish stock research." Certainly, economic theories suggest that the price constraint should facilitate the non-price competition. However, we doubt that the non-price competition will be in the form of better services and research. An increase in tick size may facilitate another round of competition in speed.

This paper is organized as follows. Section 2 describes institutional details and the data. Section 3 defines two measures of price constraint and examines their relation with high frequency trading. Section 4 examines the impact of exogenous technology shocks on speed competition, and these shocks mostly affect stocks with higher price constraint. Section 5 shows the impact of speed on liquidity and market efficiency measure based on exogenous technology shocks. Section 6 concludes the paper and discusses the policy implications.

2. Data and Institutional Details

[11] Exchanges Said to Prepare Pilot Programs for Changing Tick Sizes, Bloomberg News, June 13, 2013.

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

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

efficient construction of the limit order book using supercomputers can be found in Gai, Choi, O'Neal, Ye, and Sinkovits (2013).

Insert Table 1 about Here

Daily TAQ data provides trades and quotes for all issues traded on the NYSE, Nasdaq (OTC), and Regionals. It has more detailed information than the usual monthly TAQ data, which is widely used in academic research. A detailed comparison between daily TAQ and monthly TAQ is in Holden and Jacobsen (2013). The unique feature that facilitates our study is the millisecond time stamp, which helps to identify another technology shock that increases the speed of disseminating trading data from 3 milliseconds to 1 millisecond. This technology improvement was adopted in a staggered matter based on the alphabetic order of ticker symbols, providing a clean identification to examine the causal effect of speed on market quality.

Also, the consolidated trade file provides information on trade execution among different exchanges, which is used to calculate the market share of different trading mechanisms. This market share provides us with a nice robustness check for one of our main results. We argue that speed competition is more intense in markets with price constraint such as the NASDAQ. Price constraint leads to oversupply of liquidity and a long queue in the best ask and bid. Therefore, speed enables faster liquidity suppliers to be at the beginning of the queue with a better chance of execution. An alternative way is to supply liquidity in markets where price constraint can be bypassed. For example, EDGA, one of the two exchanges owned by Direct Edge, offers an inverted fee structure where liquidity providers need to pay a fee for executed orders. The other exchange, EDGX, has a normal fee structure where liquidity providers get the rebate while liquidity makers pay the fee. Because these two exchanges are almost identical except for the fee

structure, we use the ratio of EDGA volume to EDGX as another measure of price constraint. The ratio is computed using TAQ.

NASDAQ high-frequency data are very similar to daily TAQ in structure, but it contains identifiers to distinguish high-frequency traders from the rest. This provides us with a measure for the activity of high-frequency traders.

2.2. Identification of Technology Shocks

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

Figure 1 demonstrates the increase of speed at the matching engine level using ITCH data. Panel A demonstrates the minimum timestamp difference between two consecutive messages across the day. These two messages do not need to come from the same trader. For example, it can be the time difference between one trader's execution and another trader's cancellation. The figure shows that there is a decrease from about 950 nanoseconds to 800 nanoseconds between April 9, 2010 and April 12, 2010 and a dramatic decrease from 800 nanoseconds to 200 nanoseconds between May 21, 2010 and May 24, 2010. Panel B of Figure 1 demonstrates, for each day, the quickest execution and cancellation. As the ITCH data track the life of each individual order, we know the cancellation and execution are from the same trader.

Panel B shows that the level of the fastest cancellation and execution does not change much for the April structural break, although the volatility of the fastest cancellation and execution drastically decreases. The structural break in May, however, has a dramatic impact on latency. The fastest cancellation and execution time difference decreases from about 1.2 microseconds to between 500 and 600 nanoseconds and stays below one microsecond for all but seven days after the break. Undoubtedly, NASDAQ entered the realm of nanosecond trading after May 24, 2010. Our conversion with NASDAQ reveals that the first structural break is a consequence of the installment of the Nehalem machine engine, while the second break is more likely to originate from the high-frequency trader side.

Insert Figure 1 About Here

The outflow messages on NASDAQ-listed stocks are distributed and processed across six different channels in "unlisted trading privileges" (UTP). During our sample period, Channel 1 handles ticker symbols from A to B; Channel 2 handles ticker symbols from C to D; Channel 3 handles ticker symbols from E to I; Channel 4 handles ticker symbols from J to N; Channel 5 handles ticker symbols from O to R; and Channel 6 handles ticker symbols from S to Z. The channel assignments provide us with two clean identifications. First, because the channel assignment is unrelated to firm fundamental, an abnormal correlation for message flow for stocks in the same channel is consistent with the quote stuffing hypothesis. Second, the identified technology enhancement is adopted in a staggered way with Channel 1 being the first one upgraded. This provides us a clean identification on the causal effect of technology enhancement on market quality. Figure 2 demonstrates these staggered technology enhancements. For illustration purposes, we present a 100 millisecond snapshot of the market. Basically, before Friday, October 7, 2011, the last digits of messages in daily TAQ data for all six channels end

with 0, 3 or 7. This implies that information was broadcasted in either 3 milliseconds for time between 0 and 3 and 7 and 0 or 4 milliseconds for time between 3 and 7. Therefore, Panel A of figure 2 shows a minimum latency of 3 milliseconds of all six channels. Panel B shows that on Monday, October 10, 2011, Channel 1 was enhanced and was able to broadcast information in every millisecond. However, the other five channels were still broadcasting with 3 to 4 millisecond gaps. The same pattern continues to October 14, 2011, but on October 17, 2011, all six channels are able to broadcast information every millisecond.

Insert Figure 2 About Here

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

Insert Figure 3 About Here

Our two identification strategy has some limitations. ITCH data feed is direct, fast data feed provided by NASDAQ. The feed, however, is single thread and we do not have cross-sectional variation in technology enhancement. To address this issue, we follow the methodology of Boehmer, Saar, and Yu (2005) by using fixed effect and control for variables known for affecting market quality. Daily TAQ data provides with cleaner identification because of cross-sectional variation, but the data is from a relatively slow consolidated feed. However, most traders in the market use a consolidated data feed. High-frequency traders may subscribe to a

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

3. Price Constraint, Maker/Taker Fee and High Frequency Activity

In this session, we demonstrate the following two results. More price constraint leads to 1) more speed competition in liquidity provision and 2) higher market share in trading venues where the price constraint can be weakened by paying a liquidity maker fee. These two results also provide the economic foundation for Section 5. If speed competition is a consequence of failed price competition, an increase in speed cannot change the cost of liquidity. [26]

3.1. Measures of Price Constraint

An intuitive measure of the price constraint is the price itself. Because stocks with prices above one dollar have a tick size of one cent, the price constraint is stricter for lower-priced stocks, especially for stocks with a larger market cap. This intuition was suggested by The Wall Street Journal:

"'The lower the share price, the more attractive the stock is to high-frequency traders,' said Justin Schack, managing director at Rosenblatt Securities Inc., a brokerage and stock-

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

market research firm. 'And if you can find a stock with a low share price that is also a large-cap stock with a big float, it becomes even more attractive.'"

We also offer another measure of the price constraint: the market share of the taker/maker fee market and the maker/taker fee market. The one-cent tick size for stocks above one dollar is enforced across exchanges.[27] However, the price constraint can be bypassed. While most exchanges like NASDAQ offer rebates to liquidity providers, three exchanges, Boston, BATS-Y and EDGA, have invert fee structure that charge liquidity providers. In our sample period, we can identify the volume of EDGA and EDGX from the TAQ data, both of which are exchanges operated by Direct Edge Group. These two trading platforms have similar infrastructures and the major difference between the two platforms is their fee structures. An intuitive way to understate this measure is as follows. A larger market share for the taker/maker market implies that more liquidity can be provided at a price lower than the nominal spread, suggesting that the nominal spread is artificially high because of the price constraint.

This measure can be further justified by two seminal theoretical papers on maker/taker fee: Colliard and Foucault (2012) and Foucault, Kadan and Kendal (2013). Without tick size, Colliard and Foucault (2012) predict that a maker/taker or a taker/maker fee structure is irrelevant because the fee will be neutralized by the variations in bid-ask spread. The nominal spread of maker/taker market is lower than that of the taker/maker market, but the spreads after adjusting the fees, cum-spread, are the same in both markets. However, Rule 611 of regulation NMS prohibits trade-through, and the trade-through is defined on the nominal spread. Rule 611 states that orders should be routed to the market with the best nominal spread unless otherwise

[27] Internalized trade and trades are not subject to the sub-penny rule. However, traders who choose dark pool and internalizers may have reasons other than price constraint. Zhu (2013) and Ye (2012) provide models on the selection of trading venues across an exchange and a dark pool.

specified[28]. Therefore, under Rule 611 and infinitely small tick size, the taker/maker fee market should hardly survive because orders will be routed to the maker/taker market with a lower nominal spread. Therefore, the co-existence of taker/maker and maker/taker market is a consequence of the price constraint. Foucault, Kadan, and Kandel (2013) show that the split of the total fee between market makers and takers becomes economically meaningful when the minimum tick size restricts adjustments of the bid and ask prices. In their model, the optimal maker/taker fee is a function of tick size. To increase the trading volume and the total fee charged, stock exchange should charge liquidity makers and subsidize liquidity takers when the tick size is large and the vice versa. In reality, exchanges cannot adjust different maker/taker fee based on the characteristics of the stock. However, economic forces implies that stocks with relative large tick size will have higher volume in taker/maker fee market and stocks with relative small tick size will have more volume in maker/taker fee market. In summary, Foucault, Kadan, and Kandel (2013) analyze the theoretical impact of the tick size on volume, and we use their model and the volume of taker/maker market to back out the unobserved variable: price constraint.

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

[28] See Rule 611 (b): exceptions.

3.2 Correlations between the Two Measures of Price Constraint

The 120 stocks were constructed by Hendershott and Riordan. The sample has 40 large stocks for stocks in the 1000 largest Russell 3000 stocks, 40 medium stocks in stocks ranked from 1001-2000 in Russell 3000, and 40 medium stocks for stocks in Russell 2001-3000. Therefore, we sort the stocks 3 by 3 based on the market cap and the price level of the stock. Table 2 shows that the market share of EDGA decreases in price. For example, for large stock, the volume of EDGA is 1.5 times the volume of EDGX for low price stocks, but the ratio is only 0.7 for large stocks with the high price.

Insert Table 2 about Here

3.3. Price Constraint and High Frequency Liquidity Marking

This section provides the relationship between the price constraint and the high frequency liquidity making. High frequency trading is measured in two ways. First, the NASDAQ high frequency trading data provides the snapshot of the limit order book with the flag indicating the market depth provided by high frequency and non-high frequency traders. We use the relative market share of depth provided by high frequency traders as a measure of the first measure of high frequency liquidity provision. Second, NASDAQ high frequency data provide the information, for each trade, the maker and taker of the liquidity. There are four types of trades. NH means that a non-high frequency trader takes liquidity from a high frequency trader; HH means that a high frequency trader takes liquidity from another high frequency trader; HN means a high frequency trader takes liquidity from a non-high frequency trader, and NN means that a non-high frequency trader takes liquidity from another non-high frequency trader. Therefore, the second measure of high frequency liquidity is the percent of trades with high frequency liquidity provision, (HH+NH)/(HH+NH+HN+NN).

Insert Table 3 about Here

Table 3 demonstrates the relationship between high frequency market making with our first measure of price constraint: price. Panel A demonstrates that the percent of depth provided by high frequency trading increases monotonically with price. For large cap stocks, high frequency traders provide 33.38% of depth for high price stock, but the liquidity provision is as large as 55.62% for low price stocks. The percent of trades with high frequency liquidity providers is less than the depth provided by high frequency traders, but the results are also monotonic for large and medium cap stocks. For example, 39.60% of trades have high frequency liquidity provider for medium stocks with low price, but the number is only 22.20% for medium stock with high price. In terms of trades, small cap stocks do not have variation for high frequency liquidity provision and the level of high frequency market making is only around 19%. Small stocks generally have larger spread and the equilibrium spread is less likely to be constrained. Therefore, we observe less market share of high frequency market making in general and limited variations across different price groups. Nevertheless, the depth provided by high frequency traders still decrease monotonically with price for small stocks. The depth provided by high frequency traders are 25.12% for small cap stocks with low price, 22.61% for small stocks with medium prices and 20.75% for small stocks with high prices.

Next, we examine the relation between high frequency liquidity provisions to our second measure of price constraint, EDGA relative to EDGX market share. We sort the 120 stocks into five portfolios. Portfolio 1 have the lowest EDGA to EDGX market share and portfolio 5 has the highest EDGA to EDGX market share. Figure 4 demonstrates that the portfolio with the highest market share in taker/maker market have the highest high frequency liquidity provision.

Insert Figure 4 about Here

In terms of trade, Panel A shows that 37% of trades have a high frequency liquidity providers, and Panel B shows that 44% of depth is provided by high frequency traders. High frequency liquidity provision decreases monotonically with the market share of taker/maker fee market. For the stocks with the lowest market share in the taker/maker market, the trades with high frequency liquidity provision is only 19% and 21%, both of which are only half of the percent for stocks in the highest EDGA to EDGX market share quintile.

Finally, we run a regression with high frequency liquidity provision as dependent variable and the market share of EDGA as independent variable. The purpose of this regression is not to show causality but to show the correlation between the market share of maker/taker fee market and high frequency liquidity provision. The R-square of this regression is as high as 0.32, implying 32 percent of variation in high frequency liquidity provision can be explained by the variations in the market share of taker/maker market. Again, we are not trying to claim direct casual effect between these two variables, and they are more likely to be driven by the same underline: price constraint. The following sub-section provides discussions on the relationship between the price constraint, the maker/taker and the take/maker fee and the competition in speed.

3.4. Discussions

The previous sections reveal the following facts. First, the taker/maker fee market is more active for large stocks with low prices. Second, high frequency liquidity provision is also more active among large stocks with low prices. Third, based on the theoretical work of Colliard and Foucault (2012) and Foucault, Kadan, and Kandel (2013), we use the market share of taker/maker fee market as a proxy for price constraint, and find that stocks with larger share in taker/maker market also have more high frequency liquidity provision. Our paper is the first one

to link the literature of high frequency trading with the literature of maker/taker fee. The price constraint and the time priority shows the interdependence of these two literatures. When the price competition in the nominal spread is constrained, traders have two options. They can compete in speed to obtain the priority of execution, or pay a fee to bypass the price constraint. Time priority is also essential here. First, the time priority implies traders with higher speed have the priority in the queue. A limit order inat the front of the queue at the constrained price has a greater expected profit than other limit orders in the queue since its execution probability is higher. Second, in United States, the time priority is imposed only in each market center but not across different markets. Therefore, the taker/maker market and maker/taker market each has its own queue. This is also an important feature that drives the current landscape of competitions. With the time priority enforced across all market centers, the taker/maker market cannot exist because liquidity providers will all go to the maker/taker market to collect the fee. Time priority in each market center implies that liquidity providers can jump ahead of the queue by paying a fee in taker/maker market.

4. Technology Shock, Speed Competition and Price Constraint

Section 3 establishes the results that stocks with higher price constraint have higher high frequency liquidity provision. We further examine the impact of technology shocks on the speed competition. We establish two results. First, the technology shocks we identify dramatically increase speed. Second, the extent of the impact of the technology shocks on the speed competition depends on the price constraint. The speed competition dramatically increases for the stocks with the high price constraint. The impact for stocks with the medium price constraint is less dramatic and is not always statically significant. For the stocks with low price constraint, we can hardly see an increase in the trading speed. This section provides the connection between

the previous section (Section 3) and the next section (Section 5). On one hand, this section provides further support for results in Section 3. Section 3 finds that stocks with higher price constraint have higher high frequency liquidity provision based on static double sorting. This section provides a dynamic view of the speed competition, and we find that stocks with the higher price constraint have a larger increase in speed after the technology enhancement. On the other hand, when the speed competition is concentrated on stocks with the price constraint, it is natural that an increase in speed competition does not lead to the improvement in liquidity. Alternatively, the speed competition may simply be a consequence of the failed price competition.

Our measure of speed is based on quote life. Our dependent variable is the log of quote life for 50th, 25th, 10th, 5th, and 1st percentile deletion messages. We focus on quote lives below the median since the quote life above the median are more likely to be driven by fundamental factors such as volatility and information, whereas quote life below median are more likely to be affected by the technology capacity. Table 4 presents the result 10 days before and 10 days after the technology shock between April 9 and April 12. We run the following regression by following Park and Malinova (2013).

$$Speed_{it} = \beta_0 + \beta_1 \times after_t \times top\ constaint\ group_i + \beta_2 \times after_t \times median\ constraint\ group_i + \beta_3 \times after_t \times bottom\ constraint\ group_i + \sum firm_i + e_{it} \quad (1)$$

where $After_t$ is a dummy variable equal to 0 when t is on or before April 9, 2010 and equal to 1 when t is on or after April 12. The level of price constraint is defined over the EDGA to EDGX market share: $topconstraintgroup_i$ equals to one for the stock belonging to one third of the highest EDGA to EDGX market share group, $mediumconstraintgroup_i$ equals to one for the stock

belonging to one third of the median EDGA to EDGX market share group, and $bottomconstraintgroup_i$ equals to one for the stock belonging to one third of the bottom EDGA to EDGX market share group. $Firm_i$ is the control variable for the firm fixed effect.

The results are presented in Table 4. Because the dependent variable is defined in log value, the coefficient is interpreted as the change in percentage. Table 4 shows that the speed has dramatically decreased for stocks with the high price constraint. For this group, the medium quote life has decreased by 26% and the 10^{th} percentile quote life has decreased by 55.1%. Quote lives have decreased for stock with median price constraint across all percentiles, but the extents of the decrease are always smaller than that of the high price constraint group. As regards to the 25 percentile quote life, stocks in the medium price constraint group has a decrease of 19.4%, whereas stocks in the high price constraint group has a decrease of as large as 40.7%. For the stocks with low price constraint, the change of quote life in quote life is generally not statistically significant. To summarize, Table 4 show that after the technology shock, the speed competition becomes more intense for stocks with the high price constraint, but the impact for stocks with the low price constraint is limited.

5. Speed and Market Quality

Section 3 and 4 finds that speed competition may be a consequence of constrained price competition, which provides some intuition to understand the results in this section. When speed competition is a consequence of failed price competition, a reduction of latency cannot decrease the cost of liquidity. Speed may determine which trader provides liquidity, but different liquidity providers cannot undercut each other by price. Speed may provide a mechanism to allocate the rents created through price constraint, but it does not change the level of liquidity. The result is demonstrated using two tests. Such a division is based both on the types of each technology

shock and the econometric method. Section 5.1 introduces our measures of liquidity and market efficiency, which serves as the dependent variable in section 5.2 and 5.3. Section 5.2 demonstrates the effect of two technology shocks that reduce the latency of matching orders from microseconds to nanoseconds. Because the NASDAQ matching engine is single threaded, we have a single time series difference before and after treatment. We use the method by Boehmer, Saar, and Yu (2005) to control other variables that may affect market quality. Section 5.3 demonstrates the effect of the technology shock that reduces the latency of trading data dissemination from 3 milliseconds to 1 millisecond. This technology was first implemented in the channel that handles stocks with a ticker symbol beginning with A and B. This creates a double difference estimator for our analysis. The methodology of this part follows Hendershott, Jones, and Menkveld (2011). It is to be noticed that our technology shock is more exogenous than the one used by Hendershott, Jones, and Menkveld (2011). In Hendershott, Jones, and Menkveld (2011), the technology is enforced in a stagger matter, but the authors acknowledge that they are unsure where the sequence of enhancement is exogenous. Stocks with higher liquidity may be improved first. In our sample, selection rule is based on name, which provides a cleaner test to study the effect of technology enhancement on market quality. These two types of technology shocks reveal the same effect: an increase in speed at or below the 1 millisecond level does not affect our usual measure of liquidity and may worsen our usual measure of market efficiency.

5.1 Measure of Dependent Variables

Our measures of liquidity and price efficiency come from the ITCH data. We construct a message-by-message limit order book where the book is updated whenever there is a new message. That is, any order addition, execution or cancellation leads to a new order book. For

example, Microsoft has about 1.08 million messages on an average trading day, and we generate and store all the resulting 1.08 million order books. This provides the most accurate view of the limit order book at any point in time. The construction is implemented by the Gordon Supercomputer in San Diego Supercomputing Center.

The message-by-message order book enables us to compute a number of metrics for market quality. We calculate four measures of liquidity. Two are spread measures: the time-weighted quoted spread and the size-weighted effective spread. The other two are depth measures: the depth at the best bid and ask and the depth within 10 cents of the best bid and ask.[30] Since we construct a full limit order book, the quoted spread is measured as the difference between the best bid and ask at any time. Each quoted spread is weighted based on the life of the quoted spread to obtain the daily time-weighted quoted spread for each stock per day. The effective spread for a buy is defined as twice the difference between the trade price and the midpoint of the best bid and ask price. The effective spread for a sell is defined as twice the difference between the midpoints of the best bid and ask and the trade price. Size-weighted effective spread is defined as the size-weighted effective spread of all the trades for each stock and each day. The two depth measures, the depth at the best bid and ask and the depth within 10 cents of the best bid and ask, are weighted using the time for each stock per day.

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

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

two one-minute returns. In an efficient market, prices should approximate a random walk with no positive or negative correlation. Therefore, a ratio closer to 1 implies higher price efficiency. Finally, the daily volume measure comes from CRSP.

5.2 Shocks at Matching Engine Level: Single Time Series Difference

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

$$L_{it} = \mu_i + \alpha After_t + \beta_1 logvol_{it} + \beta_2 range_{it} + \beta_3 Prc_{it} + \varepsilon_{it}, \quad (2)$$

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

Table 5 shows that these technology shocks do not have a statistically and economically significant impact on spread. The quoted spread decreases by -0.0394 cent and the effective spread increases by 0.00115 cent, but both results are not statistically significant. The depths at the best bid and ask also do not change, but we find a 2015-share decrease of market depth within 10 cents away from the best bid and ask. Overall, we find that these two technology shocks neither increase nor decrease spread but slightly decrease the depth.

Insert Table 5 About Here

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

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

$$E_{it} = \mu_i + \lambda After_t + \varepsilon_{it}, \quad (3)$$

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

Insert Table 6 About Here

Table 6 shows that the variance ratio at the one-minute level does not have a statistically significant change before and after the technology shocks. The change in trading volume is also not statistically significant. However, volatility slightly increases after the technology shocks. The result on volatility provides additional intuition why spread does not decrease relative to the argument based on option theory. One argument that speed may improve liquidity is as follows. Limit order is the free trade option offered by liquidity providers to liquidity demanders. An increase of speed decreases duration of the option. Therefore, the price of the option decreases and liquidity improves. This argument is correct if we hold volatility as given, though the economic magnitude of the effect is not clear. When an increase of speed also increases volatility, the argument may not true. For example, in classical Black-Scholes-Merton model, the value of option depends on the product of variance and time. An increase of speed from

millisecond to nanosecond may decrease the duration of option, but it also increase the volatility. An intuitive way to understand the issue is that volatility at nanosecond should be close to zero in human world, but it can be significant when multiple computers can trade in nanoseconds. Basically, speed in can allows high frequency traders cancel their trades in nanoseconds, but it can also increases volatility because it increases possible price updates in nanoseconds. Therefore, we cannot get a conclusive answer between the relationship between speed and liquidity through option pricing models, because volatility is also a function of speed. Our empirical results suggest that the impact of speed on liquidity is insignificant.

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

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

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

$$L_{it} = \mu_i + \beta_1 pilot_t + \beta_2 after_t + \beta_3 pilot_t after_t * + \varepsilon_{it} \quad (4)$$

μ_i is the stock random effect. L_{it} is the daily liquidity measure such as time-weighted quoted spread, size-weighted effective spread, time-weighted depth at the best bid and ask and time-weighted depth within 10 cents of the best bid and ask. The variable of interest is the diff-in-diff estimator β_3. The result is summarized in Table 7. The first two columns are the effect of the first technology enhancement, and the third and fourth columns summarize the effect of the second technology enhancement. We find similar effects in the previous section: these technology shocks do not improve liquidity.

Insert Table 7 about Here

6. Conclusion

This paper provides a unified empirical explanation for two important features in current market: the competition in speed and the competition between maker/taker and taker/maker markets. We argue that two deviations from Walrasian equilibrium, discrete price and time priority, are the economic forces that drive these competitions.

Because rule 612 of regulation NMS prohibits price competitions with increment less than one cent, speed allocates executions when the equilibrium bid-ask spread is within two ticks. Therefore, we observe the speed competition in the liquidity provision is more intense for large stocks with low prices, which in turn have larger relative tick sizes. These stocks also have a larger market share in the taker/maker market, where the price constraint can be bypassed by paying a fee. Based on the theoretical model of Colliard and Foucault (2012) and Foucault,

Kadan, and Kandel (2013), we use the market share of the taker/maker market relative to the maker/taker market as our second measure of price constraint besides price. We find that high frequency liquidity provision is concentrated in stocks with high price constraint, implying speed competition is more intense when the price competition is constrained.

In addition, we not only find that stocks with price constraint have relatively higher speed of trading, but also find that they have relatively larger increase in speed during technology shocks. On the contrary, the trading speed of stocks with low price constraint hardly increases at all. This evidence, again, suggests that the speed competition should be a consequence of the constrained price competition.

Not surprisingly, if the speed competition is a consequence of the constrained price competition, an increase in speed will not impact liquidity. We find that two specific technology shocks, which exogenously increase the speed of matching orders from the microsecond level to the nanosecond level. However, the spread does not decrease following the technology shock and the variance ratio does not improve. On the contrary, market depth decreases and short-term volatility increases. We also explore the technology shocks that increase the speed to disseminate trading data. These shocks happened in a stagger manner based on the alphabetic order of stock names, which provides ideal randomness for the diff-in-diff test. Again, we find similar results: an increase in speed does not improve liquidity.

The policy makers have strong concern over the speed competition. Currently, there are several proposed policies such as imposing a minimum quote life, charging cancellation fees and transaction taxes. The finance literature before the sub-millisecond era, on the other hand, shows that an increase in speed benefits liquidity. Our results supported by economic reasoning provide a relatively neutral view on the high frequency trading: high frequency trading neither improves

nor harms liquidity. Regulation 612, which constrains the price competition, together with the time priority rule, encourage the speed competition at the constrained price. Our results suggest that the speed competition above certain threshold is a zero-sum game, the existence of which is facilitated by the rent created through the price constraint. The price constraint also facilitates the co-existence and competition for markets with different fee structures.

Therefore, in the debate on whether and how should we regulate high frequency traders, we should first be aware of the fact that high frequency trading may be a consequence of existing regulations. Therefore, if policy makers have concerns on the speed competition, one solution is to decrease the tick size, which will enforce competition to focus more on price. Interestingly, from an economic perspective, this is the deregulation instead of the regulation, and this will move the market closer to a Walrasian equilibrium. The other solution is to decrease the importance of time priority below the millisecond level, where orders that arrive at the same millisecond share the same time priority. This policy is different from imposing a minimum quote life, thought it slows down the market as well. Minimum quote life policy will change the relative speed of liquidity providers relative to liquidity demanders, and the consequence of such a policy is ambiguous. Decreasing the important of time priority, however, affects all traders in a similar manner.

Certainly, the above two policies need to be justified by both implementation benefits and costs. We are concerned, however, about a recent reverse trend. According to Bloomberg, SEC is considering increasing the tick size for small stocks, and the reason is that "[m]arket makers and dealers need more economic incentives to bring smaller companies public, provide bids and offers and publish stock research." The argument that the price constraint leads to the non-price competition is generally true, but we doubt whether the non-price competition will be in the form

of publishing stock research. A more direct way for the non-price competition is the competition in speed. Our results suggest that an increase in tick size can facilitate another round of the speed competition.

Our paper is the first empirical paper to define the price constraint and its impact on high frequency trading, maker/taker and taker/maker fee markets, and it can be extended in different ways. First, the current theoretical work on the speed competition focuses on the role of information. Our paper points out another channel of the speed competition: price constraint. Models with discrete prices can be constructed to understand the value of speed and the impact of the price constraint on the market quality. Second, we explain the market share of taker/maker fee based on the price constraint, and theoretical models can be built to understand why there exists separating equilibrium for traders on different trading platforms, and how exchanges set fee structures in both markets to maximize the total profits. Empirically, the relation between price constraint, speed competition and the maker/taker versus taker/maker fees can be further explored, probably by using better data. For example, SEC recently announce a pilot program to increase the tick size for a number of small stocks, it would be interesting to see the impact of this shock on the speed competition.

REFERENCES

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

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

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

Brolley, Michael, and Katya Malinova, 2012, Informed trading and maker-taker fees ina low-latency limit order market, Working paper, University of Toronto.

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

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

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

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

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

Colliard, Jean-Edouard and Foucault Thierry, 2012, Trading fees and efficiency in limit order markets" *Review of Financial Studies*, 25, 3389-3421.

Foucault, Thierry, Ohad Kadan and Eugene Kandel, 2013 Liquidity cycles, and make/take fees in electronic market, *Journal of Finance* 68, 299-341.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Malinova, Katya and Andreas Park, Subsidizing liquidity: the Impact of make/take fees on market quality, forthcoming in *Journal of Finance*

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

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

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

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

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

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

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

Table 2: Consistency of Price Constraint Measures

This table demonstrates the ratio of Direct Edge A (EDGA) volume to Direct Edge X (EDGX). EDGA charges liquidity providers and rebates liquidity takers whereas EDGX charges liquidity takers and rebates liquidity makers. The sample includes 120 stocks in NASDAQ high-frequency data. The stocks are first sorted on market cap and then on price.

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

Table 3: High Frequency Trading and Price

This table demonstrates the market share of high-frequency liquidity provision. Panel A demonstrates the depth provided by high frequency traders whereas Panel B demonstrates the share of trades with high frequency liquidity providers. The 120 stocks in NASDAQ high-frequency data are first sorted on market cap and then sorted on price.

Panel A: Percentage of BBO Depth Provided by High-Frequency Traders			
	Low Price	Medium Price	High Price
Large Cap	55.62%	46.17%	33.38%
Medium Cap	39.60%	30.26%	24.55%
Small Cap	25.12%	22.61%	20.75%
Panel B: Percentage of Trading Provided by High-Frequency Traders			
	Low Price	Medium Price	High Price
Large Cap	45.40%	37.50%	30.90%
Medium Cap	35.50%	23.40%	22.20%
Small Cap	18.60%	18.90%	18.50%

Table 4: Technology Shock, Price Constraint and Speed

The table examines the impact of technology shocks on speed competition for stocks with different level of price constraint. The dependent variables ($speed_{it}$) are the logs of the 50th, 25th, 10th, 5th, and 1st percentile deletion message quote lives for each stock. The technology enhancement occurred between April 9 and April 12. The estimation is based on the following regression specification:

$$Speed_{it} = \beta_0 + \beta_1 \times after_t \times topconstaint\ group_i + \beta_2 \times after_t \times medium\ constraint\ group_i + \beta_3 \times after_t \times bottomconstraint\ group_i + \sum firm_i + e_{it}$$

$after_t$ is 1 for 10 trading days after the technology shock and $after_t$ is 0 for 10 trading days before the technology shock. *top constraint group*$_i$, *medium constraint group*$_i$, and *bottom constraint group*$_i$ are dummies that are 1 when *stock*$_i$ is in the respective group with respect to the ratio of Direct Edge A and Direct Edge X trading volume. Standard errors are in parentheses. Coefficients for firm fixed effect variables are not reported for brevity. * indicates significance at the 10% level,** at the 5% level, and *** at the 1% level.

VARIABLES	(1) logp50	(2) logp25	(3) logp10	(4) logp5	(5) logp1
$after_t \times top\ third\ constraint\ group_i$	-0.260***	-0.407***	-0.551***	-0.366***	-0.200***
	(0.0723)	(0.0965)	(0.0940)	(0.0720)	(0.0390)
$after_t \times medium\ third\ constraint\ group_i$	-0.230***	-0.194*	-0.224**	-0.111	-0.125***
	(0.0743)	(0.0991)	(0.0966)	(0.0739)	(0.0400)
$after_t \times bottom\ third\ constraint\ group_i$	0.00731	0.162*	0.111	0.0786	-0.0463
	(0.0733)	(0.0978)	(0.0953)	(0.0730)	(0.0395)
Constant	2.388***	-0.314***	-3.507***	-5.509***	-7.583***
	(0.0299)	(0.0399)	(0.0388)	(0.0297)	(0.0161)
Observations	2,285	2,285	2,285	2,285	2,285
R-squared	0.010	0.011	0.019	0.013	0.017

Table 5: The Impact of Speed Improvement of Submitting or Processing Orders on Liquidity

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

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

Table 6: The Impact of Speed Improvement of Submitting or Processing Orders on Market Efficiency and Volume

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

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

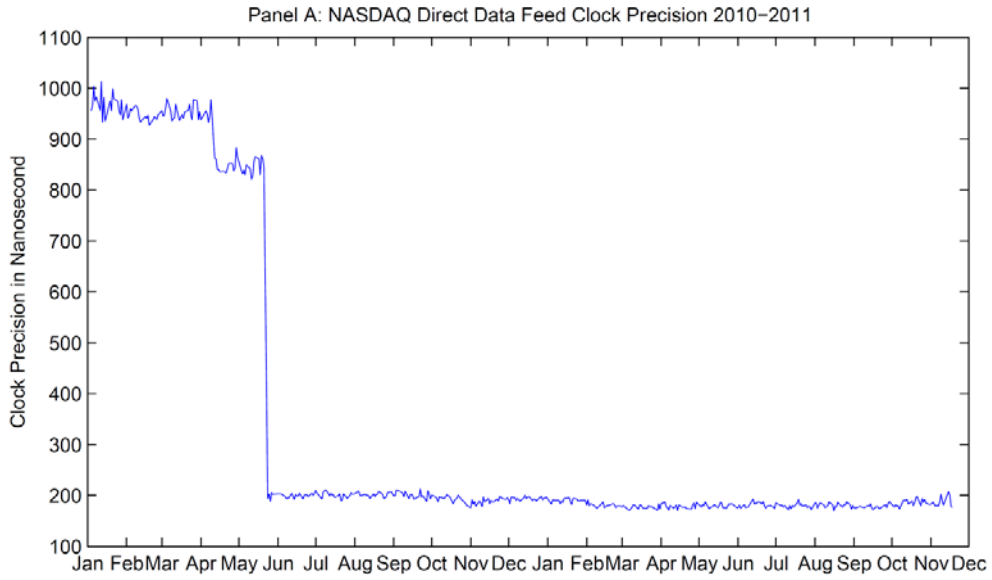
Table 7: The Impact of Speed Improvement of the Consolidated Tape on Liquidity

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

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

Figure 1: Technology Shocks on Submitting and Cancelling Orders

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



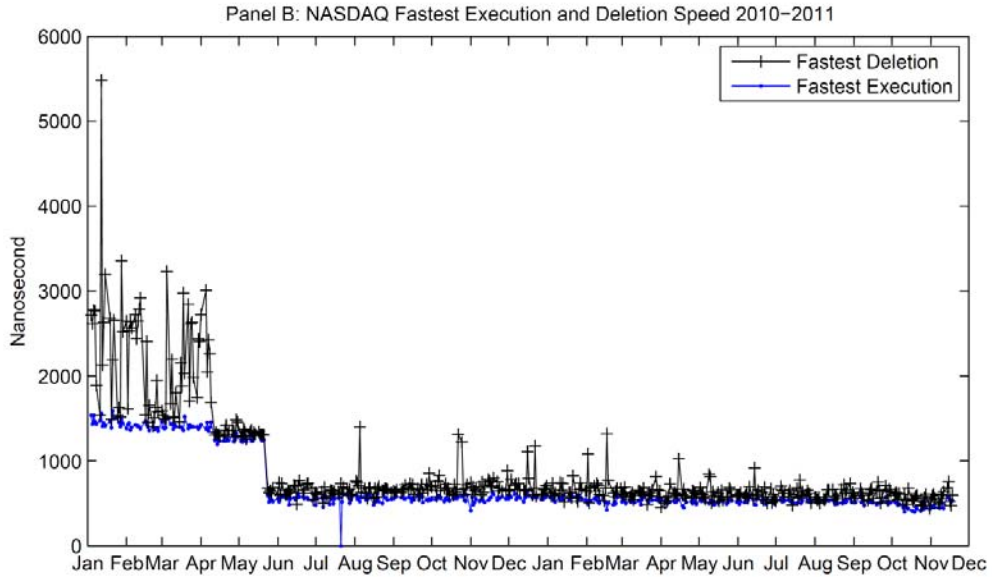


Figure 2: Technology Shocks on Transmitting Trading Data

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

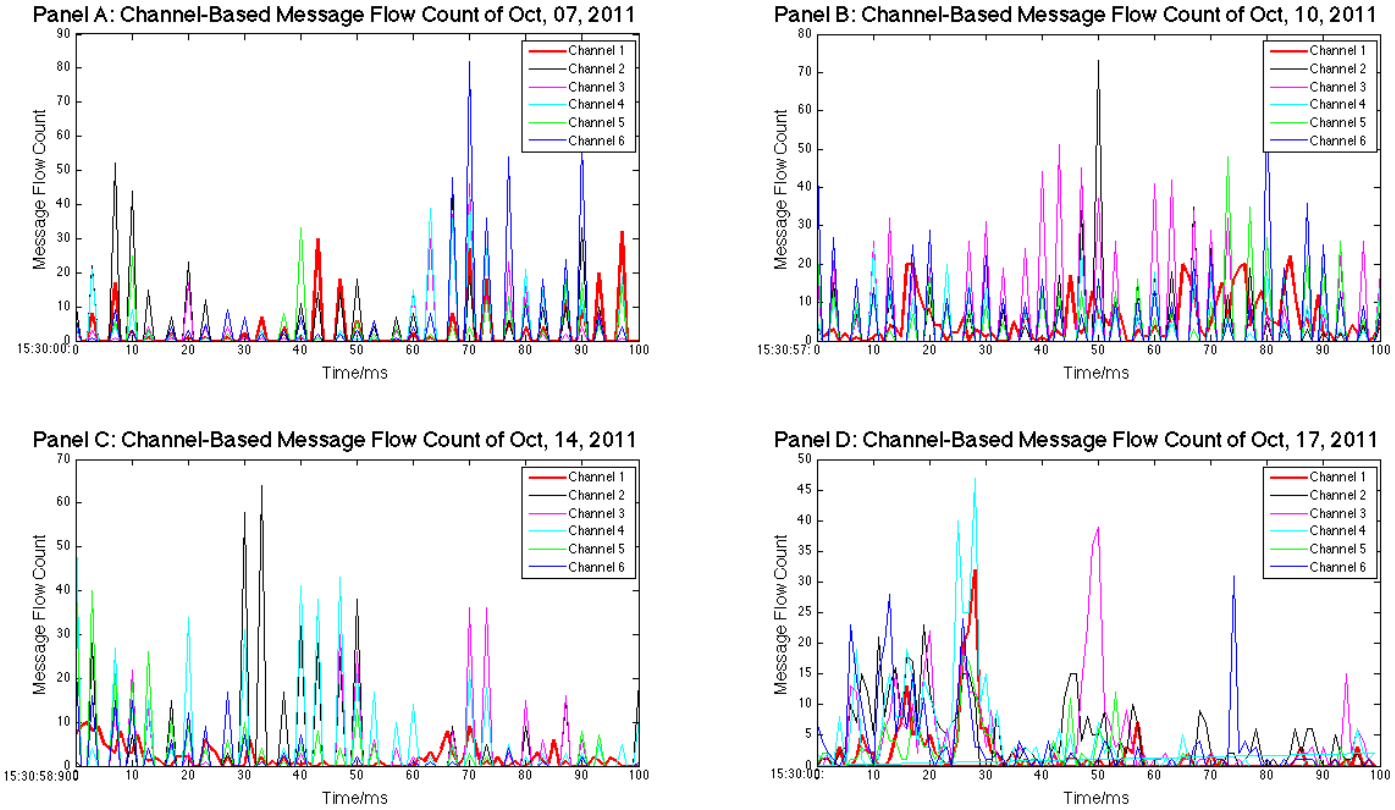


Figure 3: Median Time Gaps between Two Broadcasts

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

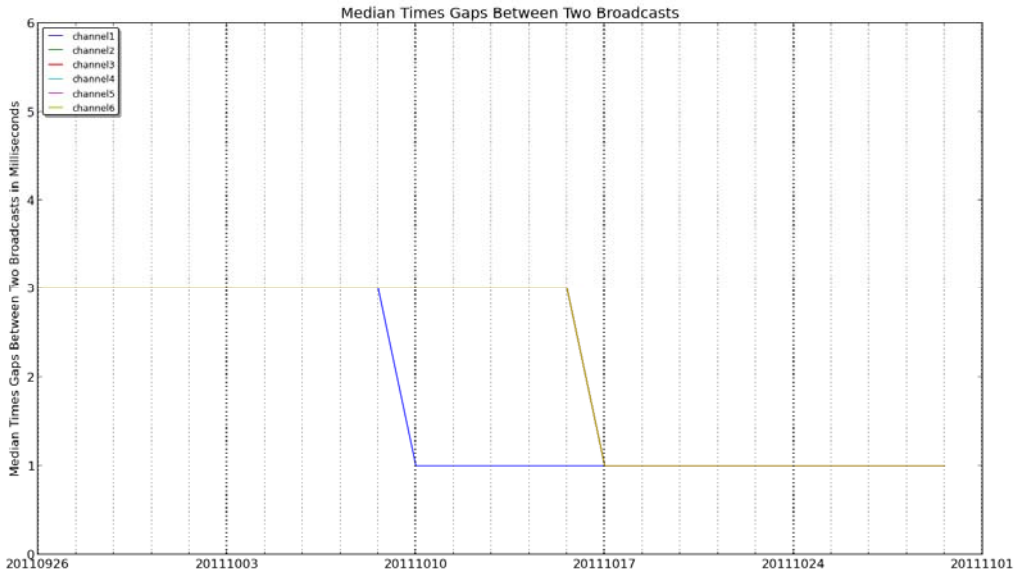
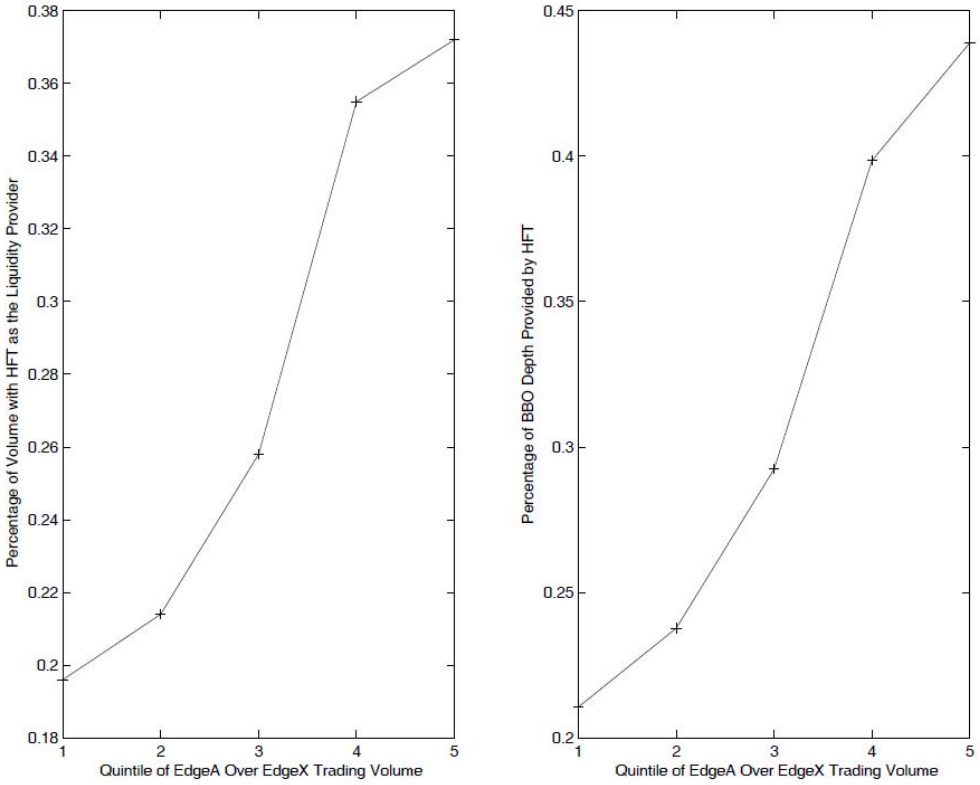


Figure 4: High Frequency Liquidity Provision and Share of Taker/Maker Fee Market

This figure relates high frequency liquidity provision with our second measure of price constraint: the volume of EDGA relative to EDGX. Stocks in quintile 1 have the lowest level of volume in EDGA (taker/maker fee market) to EDGX (maker/taker fee market) and have the lowest price constraint. Stocks in quintile 5 have the highest price constraint with the largest market share in taker/maker market. Panel A presents the percent of volume with high frequency liquidity providers and panel B provides the percent of depth provided by high frequency traders.



Appendix: Structure of ITCH Data

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

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

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

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

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

which is known from the U message, and the stock and the buy/sell indicator can be found at the A message.