# Low-Latency Trading

Joel Hasbrouck and Gideon Saar

This version: May 2011

Joel Hasbrouck is from the Stern School of Business, 44 West 4th Street, New York, NY 10012 (Tel: 212-998-0310, jhasbrou@stern.nyu.edu). Gideon Saar is from the Johnson Graduate School of Management, Cornell University, 455 Sage Hall, Ithaca, NY 14853 (Tel: 607-255-7484, gs25@cornell.edu). We are grateful for comments from Andrew Karolyi, Albert Menkveld, Ciamac Moallemi, Maureen O'Hara, and seminar (or conference) participants at Cornell's Johnson School, Cornell Financial Engineering Manhattan, the CREATES Market Microstructure Symposium (Aarhus), ESSEC Business School, Humbolt University, the National Bureau of Economic Research Market Microstructure Group meeting, New York University, the Chicago Quantitative Alliance / Society of Quantitative Analysts, the Investment Industry Regulatory Organization of Canada / DeGroote School, Rutgers Business School, and the World Federation of Stock Exchanges Statistics Advisory Group.

# Low-Latency Trading


## Abstract

This paper studies market activity in the "millisecond environment," where computer algorithms respond to each other almost instantaneously. Using order-level NASDAQ data, we find that the millisecond environment consists of activity by some traders who respond to market events (like changes in the limit order book) within roughly 2-3 ms, and others who seem to cycle in wall-clock time (e.g. access the market every second). We define low-latency activity as strategies that respond to market events in the millisecond environment, the hallmark of proprietary trading by a new breed of high-frequency traders We construct a measure of low-latency activity by identifying "strategic runs," which are linked submissions, cancellations, and executions that are likely to be parts of a dynamic strategy. We use this measure to study the impact that low-latency activity has on market quality both during normal market conditions and during a period of declining prices and heightened economic uncertainty. Our conclusion is that increased low-latency activity improves traditional market quality measures such as short-term volatility, spreads, and displayed depth in the limit order book.

## I. Introduction

Our financial environment is characterized by an ever increasing pace of both information gathering and the actions prompted by this information. Speed is important to traders in financial markets for two main reasons. First, the inherent fundamental volatility of financial securities means that rebalancing positions faster could result in higher utility. Second, irrespective of the absolute speed, being faster than other traders can create profit opportunities by enabling a prompt response to news or market-generated events. This latter consideration appears to drive an arms race where traders employ cutting-edge technology and locate computers in close proximity to the trading venue in order to reduce the latency of their orders and gain an advantage. As a result, today's markets experience intense activity in the "millisecond environment," where computer algorithms respond to each other at a pace 100 times faster than it would take for a human trader to blink.

While there are many definitions for the term "latency," we view it as the time it takes to learn about an event (e.g., a change in the bid), generate a response, and have the exchange act on the response.[1] Exchanges have been investing heavily in upgrading their systems to reduce the time it takes to send information to customers as well as to accept and handle customers' orders. They have also begun to offer traders the ability to co-locate the traders' computer systems next to theirs, thereby reducing transmission times to under a millisecond (a thousandth of a second). As traders have also invested in the technology to process information faster, the entire event/analysis/action cycle has been reduced for some traders to a few milliseconds.

An important question is, who benefits from such massive investment in technology? After all, most trading is a zero sum game, and the reduction in fundamental

[1] More specifically, we define latency as the sum of three components: the time it takes for information to reach the trader, the time it takes for the trader's algorithms to analyze the information, and the time it takes for the generated action to reach the exchange and get implemented. The latencies claimed by many trading venues, however, are usually defined much more narrowly, typically as the processing delay measured from the entry of the order (at the vendor's computer) to the transmission of an acknowledgement (from the vendor's computer).

risk mentioned above would seem very small for time intervals on the order of several milliseconds. There is a new breed of high-frequency traders in the market who implement low-latency strategies, which we define as strategies that respond to market events in the millisecond environment. These traders now generate most message activity in financial markets and according to some accounts also take part in the majority of the trades.[2] While it appears that intermediated trading is on the rise (with these low-latency traders providing liquidity to other market participants), it is unclear whether intense low-latency activity harms or helps market quality.

Our goal in this paper is to examine the influence of these low-latency traders on the market environment. We begin by studying the millisecond environment to ascertain how low-latency strategies affect the time-series properties of market activity. We then ask the following question: how does the interaction of these traders in the millisecond environment impact the quality of markets that human investors can observe? In other words, we would like to know how their combined activity affects attributes such as the short-term volatility of stocks, the total price impact of trades, and the depth of the market. To investigate these questions, we utilize NASDAQ order-level data (TotalView-ITCH) that are identical to those supplied to subscribers and which provide real-time information about orders and executions on the NASDAQ system. Each entry (submission, cancellation, or execution of an order) is time-stamped to the millisecond, and hence these data provide a very detailed view of activity on the NASDAQ system.

We find that the millisecond environment shows evidence of two types of activities: one by traders who seem to operate according to a schedule (e.g., access the market every second) and the other by traders who respond to market events. The former is likely generated by agency algorithms employed to minimize trading costs of buy-side managers, and it creates periodicities in the time-series properties of market activity based on wall-clock time. In contrast, we believe that strategies that respond to market events (i.e., low-latency activity) is the hallmark of proprietary trading by a new set of

[2] See, for example, the discussion of high-frequency traders in the SEC's Concept Release on Equity Market Structure.

proprietary high-frequency traders that feature prominently in today's market environment.

We use the data to construct "strategic runs" of linked messages that describe dynamic order placement strategies. By tracking submissions, cancellations, and executions that can be associated with each other, we create a measure of low-latency activity. We use a simultaneous equation framework to examine how the intensity of low-latency activity affects market quality measures. We find that an increase in low-latency activity lowers short-term volatility, reduces quoted spreads and the total price impact of trades, and increases depth in the limit order book. If our econometric framework successfully corrects for the simultaneity between low-latency activity and market attributes, then increased activity of low-latency traders in the current market environment is beneficial to the traditional benchmarks of market quality.

Furthermore, we employ two distinct sample periods to investigate whether the impact of low-latency trading on market quality (and the millisecond environment in general) differs between calm days and periods of declining prices and heightened uncertainty. Over October 2007, our first sample period, stock prices are relatively flat or slightly increasing. Over our second sample period, June 2008, stock prices are declining (the NASDAQ index is down 8% in that month) and uncertainty is high following the fire sale of Bear Stearns. We find that the millisecond environment with its various attributes is rather similar across the two sample periods. More importantly, higher low-latency activity enhances market quality in both environments, and is especially beneficial in reducing volatility for small stocks during stressful times.[3]

Our paper relates to the small but growing strands in the literature on speed in financial markets and algorithmic trading. In particular, Riordan and Storkenmaier (2008), Easley, Hendershott, and Ramadorai (2009), and Hendershott and Moulton (2009) examine market-wide changes in technology that reduce the latency of

[3] We note that this does not imply that the activity of low-latency traders would help curb volatility during extremely brief episodes such as the "flash crash" of May 2010, in which the market declined by about 7% over a 15-minute interval before partially rebounding.

information transmission and execution, but reach conflicting conclusions as to the impact of such changes on market quality. There are several papers on algorithmic trading that characterize the trading environment on the Deutsche Boerse (Gsell (2008), Gsell and Gomber (2008), Groth (2009), Prix, Loistl, and Huetl (2007), Hendershott and Riordan (2009)), and two papers that study U.S. markets: Hendershott, Jones, and Menkveld (2009) and Brogaard (2010). None of these papers study the characteristics of the millisecond environment, but the latter two papers attempt to evaluate the impact of algorithmic trading on market quality in the U.S., a goal we share as well.[4]

The rest of this paper proceeds as follows. The next section describes our sample and data. Section III characterizes the new trading environment. We provide evidence on intensity and periodicity in the millisecond environment, and construct a measure of low-latency activity designed to capture dynamic strategies. Section IV studies how the activity of low-latency traders in the millisecond environment influences traditional attributes of market quality such as liquidity and short-term volatility. In Section V we discuss related papers and place our findings within the context of the literature. Section VI concludes the paper with a discussion of low-latency trading from economic and regulatory perspectives.

## II. Data and Sample

### *II.A. NASDAQ Order-Level Data*

The NASDAQ Stock Market is a pure agency market. It operates an electronic limit order book that utilizes the INET architecture (which was purchased by NASDAQ in 2005).[5] All submitted orders must be price-contingent (i.e., limit orders), and traders who seek immediate execution need to price the limit orders to be marketable (e.g., a buy order priced at or above the prevailing ask price). Traders can designate their orders to display in the NASDAQ book or mark them as "non-displayed," in which case they reside in the

[4] The joint CFTC/SEC report on the "flash crash" of May 6, 2010, looks at the role of high-frequency trading in this extreme episode (U. S. Commodity Futures Trading Commission and the U.S. Securities and Exchange Commission, 2010). Although much can be learned from extreme events, our study, in contrast, uses sample periods that are longer and arguably more representative.

[5] See Hasbrouck and Saar (2009) for a more detailed description of the INET market structure.

book but are invisible to all traders. Execution priority follows price, visibility, and time. All displayed quantities at a price are executed before non-displayed quantities at that price can trade.

The NASDAQ data we use, TotalView-ITCH, are identical to those supplied to subscribers, providing real-time information about orders and executions on the NASDAQ system. These data are comprised of time-sequenced messages that describe the history of trade and book activity. Each message is time-stamped to the millisecond, and hence these data provide a detailed picture of the trading process and the state of the NASDAQ book. We are able to observe four different types of messages in the TotalView-ITCH dataset: (i) the addition of a displayed order to the book, (ii) the cancellation of a displayed order, (iii) the execution of a displayed order, and (iv) the execution of a non-displayed order.

With respect to executions, we believe that the meaningful economic event is the arrival of the marketable order. In the data, when an incoming order executes against multiple standing orders in the book, separate messages are generated for each standing order. We view these as a single marketable order arrival, so we group as one event multiple execution messages that have the same millisecond time stamp, are in the same direction, and occur in a sequence unbroken by any non-execution message. The component executions need not occur at the same price, and some (or all) of the executions may occur against non-displayed quantities.

### *II.B. Sample*

Our sample is constructed to capture variation across firms and across market conditions. We begin by identifying all common, domestic stocks in CRSP that are NASDAQ-listed in the last quarter of 2007.[6] We then take the top 500 stocks, ranked by market capitalization as of September 30, 2007. Our first sample period is October of 2007 (23

[6] NASDASQ introduced the three-tier initiative for listed stocks in July of 2006. We use CRSP's NMSIND=5 and NMSIND=6 codes to identify eligible NASDAQ stocks for the sample (which is roughly equivalent to the former designation of "NASDAQ National Market" stocks).

trading days). The market was relatively flat during that time, with the S&P 500 Index starting the month at 1,547.04 and ending it at 1549.38. The NASDAQ Composite Index was relatively flat but ended the month up 4.34%. Our October 2007 sample is intended to reflect a "normal" market environment.

Our second sample period is June 2008 (21 trading days), which represents a period of heightened uncertainty in the market, falling between the fire sale of Bear Stearns in March of 2008 and the Chapter 11 filing of Lehman Brothers in September. During June, the S&P 500 Index lost 7.58%, and the NASDAQ Composite Index was down 7.99%. In this sample, we continue to follow the firms used in the October 2007 sample, less 29 stocks that were acquired or switched primary listing. For brevity, we refer to the October 2007 and June 2008 samples as "2007" and "2008," respectively.

In our dynamic analysis we use summary statistics constructed over 10-minute intervals. To ensure the accuracy of these statistics, we impose a minimum message count cutoff. A firm is excluded from a sample if more than ten percent of the 10-minute intervals had fewer than 250 messages. Google and Apple are excluded due to computational limitations. Net of these exclusions, the 2007 sample contains 345 stocks, and the 2008 sample contains 394 stocks.

Table 1 provides summary statistics for the stocks in both sample periods using information from CRSP and the NASDAQ dataset. Panel A summarizes the measures obtained from CRSP. In the 2007 sample, market capitalization ranges from $789 Million to $276 Billion, with a median of slightly over $2 Billion. The sample also spans a range of trading activity and price levels. The most active stock exhibits an average daily volume of 77 million shares; the median is about one million shares. Average closing prices range from $2 to $272 with a median of $29. Panel B summarizes data collected from NASDAQ. In 2007 the median firm had 27,130 limit order submissions (daily average), 24,374 limit order cancellations, and 2,489 marketable order executions.[7] Statistics for the 2008 sample are similar.

[7] These counts reflect our execution grouping procedure. In 2007, for example, the mean number of order submissions less the mean number of order cancellations implies that the mean number of executed

## III. Characterizing the New Trading Environment

### *III.A. Intensity, periodicity, and the Speed of Response to Market Events*

Current market observers often comment on the rapid pace of activity. In fact, the typical average message rate is unremarkable. The sum of the median numbers of limit order submissions, cancellations, and marketable order executions for 2007 is 53,993. With 23,400 seconds in a 6.5 hour trading session, a representative average message arrival rate is about 2.3 messages per second.

The average, however, belies the intensely episodic nature of the activity. To illustrate this, we estimate the hazard rate for the inter-message durations. The hazard rate is the message arrival intensity (for a given stock), conditional on the time elapsed since the last message (for that stock). Figure 1 depicts graphs of the hazard functions for two types of messages: (i) those that do not involve the execution of trades (arrivals and cancellations of nonmarketable limit orders), and (ii) executions of trades (against both displayed and non-displayed limit orders). Panel A presents the hazard rates up to 100 ms, while Panel B shows the hazard rates up to 1000 ms (i.e., one second). The hazard rates we observe in the market exhibit three striking characteristics: a very high initial level, a rapid decline, and (in the case of non-execution events) a small number of apparent peaks.

In the first millisecond (after the preceding message) the hazard rate for submissions/cancellations is 334 messages per second in 2007, and 283 messages per second in 2008, i.e., roughly one hundred times the average arrival intensity. These high values, however, rapidly dissipate. In 2008, the initial hazard rate drops by about 90 percent in the first ten milliseconds, and by about 98% in the first hundred milliseconds.

A declining hazard rate is consistent with event clustering. This is a common feature of financial data, and is often modeled statistically by dependent duration models (e.g., Engle and Russell (1998), and Hautsch (2004)). From an economic perspective,

---

standing limit orders is 41,447–37,126=4,321. This is above the reported mean number of marketable orders executed (3,593) because a single marketable order may involve multiple standing limit orders. As we describe in Section II.A, we group executions of standing limit orders that were triggered by a single marketable order into one event.

variation in trading intensity has long been believed to reflect variation in information intensity. While the information can be diverse in type and origin, it is often viewed as relating to the fundamental value of the stock and originating from outside the market (e.g., a news conference with the CEO or a change in an analyst's earnings forecast). At horizons of extreme brevity, however, there is simply not sufficient time for an agent to be reacting to anything *except* very local market information.[8] The information is about whether someone is interested in buying or selling, and it may lead to a transient price movement rather than a permanent shift.

While the hazard rate graphs are dominated by the rapid decay, they also exhibit local peaks. Over the very short run (Panel A), submissions/cancellations have distinct peaks in both the 2007 and 2008 samples at around 60 ms. The magnitude of the peaks is rather large. For example, the peak at around 60 ms in the 2007 sample implies a hazard rate that is twice as large as the hazard rate one would get by averaging the rates a few milliseconds before and after this specific duration. There are also discernible peaks at 11-12 ms. These are somewhat less visible because they occur in a region dominated by the rapid decay. They are nevertheless about 30% higher than the average surrounding values. These peaks do not appear as distinctly in the execution hazard rates. The latter, however, also peak around 2-3 ms, a feature discussed in more detail below. Over a longer interval (Panel B), submissions/cancellations exhibit peaks around 100 and (partially visible) 1,000 ms.

The peaks at 60, 100 and 1,000 ms may reflect algorithms that access the market periodically. To further characterize the periodicities, we examine the level of activity in wall-clock time (the hazard rate analyses are effectively set in event time). The timestamps in the data are milliseconds past midnight. Therefore for a given timestamp *t*, the quantity $\mathrm{mod}(t,1000)$ is the millisecond remainder, i.e., a millisecond time stamp

[8] It is unlikely that the time it takes to process and extract the pricing-relevant implications of fundamental information (e.g., statements made by the CEO of a firm) is as low as 2-3 ms. Furthermore, the frequency of fundamental information events is so low that orders reacting to such events are unlikely to generate observable peaks in the hazard rates that are computed from tens of thousands of observations for each stock (in one month).

within the second. Assuming that message arrival rates are constant or (if stochastic) well-mixed within a sample we would expect the millisecond remainders to be uniformly distributed over the integers $\{0,1,\ldots,999\}$.

The data, however, tell a different story. Panel A of Figure 2 depicts the sample distribution of the millisecond remainders. The null hypothesis is indicated by the horizontal line at 0.001. The distributions in both sample periods exhibit marked departures from uniformity. Both feature large peaks occurring shortly after the one-second boundary (at roughly 10-30 ms), and also around 150 ms. Broad elevations occur around 600 ms. We believe that these peaks are indicative of automated trading systems that periodically access the market, near the second and the half-second. These intervals are substantially longer than the sub-100 ms horizon that characterizes the elevated hazard rates.

In other words, unlike low-latency traders who respond to market-created events, these algorithms submit an order and periodically revisit it. These periodic checks would also be subject to latency delays. For example, even if an algorithm is programmed to revisit an order exactly on the second boundary, any response would occur subsequently. The time elapsed from the one-second mark would depend on the latency of algorithm (i.e., how fast the algorithm receives information from the market, analyzes it, and responds by sending messages to the market). The observed peaks at 10-30 ms or at 150 ms could be generated by clustering in transmission time (due to geographic clustering of algorithmic trading firms), technology, or simply the large volume handled by particular firms.

To investigate whether there might exist longer periodicities, we construct the sample distribution of timestamps mod 10,000 (Figure 2, Panel B). These graphs are dominated by the strong one-second cycles, but also appear to contain two- and ten-second variations.

Note that the peaks observed at shorter durations in Figure 1 (e.g., 2-3 or 11-12 ms), may represent strategic responses to market events rather than periodicity, and so serve as useful indications of effective latency. Our definition of low-latency trading is

"strategies that respond to market events in the millisecond environment." Although any event might be expected to affect all subsequent events (as in Figure 1), our interest in the speed of response suggests focusing on conditioning events that seem especially likely to trigger rapid reactions. One such event is the improvement of a quote. An increase in the bid may lead to an immediate trade (against the bid) as potential sellers race to hit it. Alternatively, competing buyers may race to cancel and resubmit their own bids to remain competitive and achieve or maintain time priority. We call the former response a same-side execution, and the latter response a same-side submission/cancellation. Events on the sell side of the book, subsequent to a decrease in the ask price, are defined similarly.

Our analysis requires only a slight change to the estimation of the hazard rates depicted in Figure 1. These earlier results are unconditional in the sense that they reflect durations subsequent to events of all types. The present characterization focuses on hazard rates subsequent to order submissions that improve the quote. Figure 3 (Panel A) depicts the conditional hazard rates for same-side events (pooled over bid increases and ask decreases).

The small local peaks at approximately 2-3 ms that we identified in Figure 1 are much more sharply defined in the conditional analysis of Figure 3, particularly for executions. This suggests that the fastest responders are subject to 2-3 ms latency. For comparison purposes, we note that human reaction times are generally thought to be on the order of 200 milliseconds (Kosinski (2010)). It is therefore reasonable to assume that these responses represent actions by automated agents (various types of trading algorithms). The figure suggests that the time it takes for some low-latency traders to observe the market event, process the information, and act on it is indeed very short.

The hazard rates depicted in Panel B of Figure 3 are conditional on an order cancellation that resulted in the deterioration of the quote (a drop in the bid or increase in the ask). Peaks at 2-3 ms are visible for same-side submissions and cancellations, presumably reflecting the repricing of orders. For executions, the peak is very small in

2007 and non-existent in 2008. Perhaps unsurprisingly, withdrawal of a bid (for example) does not induce sellers to chase it.

The millisecond environment therefore consists of activity by some traders who respond to market events and others who seem to cycle in wall-clock time. Before we proceed to measure low-latency trading and investigate its impact on market quality, it is useful to discuss the types of market participants whose activities shape the millisecond environment.

### *III.B. The Players: Proprietary Algorithms and Agency Algorithms*

Much trading and message activity in U.S. equity markets is commonly attributed to trading algorithms.[9] However, not all algorithms serve the same purpose and therefore the patterns they induce in market data and the impact they have on market quality could depend on their specific objectives. Broadly speaking, however, we can categorize algorithmic activity as agency or proprietary. Agency algorithms are used by buy-side institutions to minimize the cost of executing trades in the process of implementing changes in their investment portfolios. While proprietary algorithms can be used for various purposes, our focus is on algorithms employed by hedge funds, proprietary trading desks of large financial firms, and independent specialty firms that are meant to profit from the trading environment itself (as opposed to investing in stocks).

Agency Algorithms are used by buy-side institutions and the brokers who serve them to buy and sell shares. They have been in existence for about two decades, but the last ten years have witnessed a dramatic increase in their appeal due to decimalization (in 2001) and increased fragmentation in U.S. equity markets (following Reg ATS in 1998 and Reg NMS in 2005). These algorithms break up large orders into pieces that are then sent over time to multiple trading venues.[10] The algorithms determine the size, timing,

[9] The SEC's Concept Release on Equity Market Structure cites media reports that attribute 50% or more of equity market volume to proprietary "high-frequency traders." A report by the Tabb Group (July 14, 2010) suggests that buy-side institutions use "low-touch" agency algorithms for about a third of their trading needs.

[10] See, for example, Bergan and Devine (2005).

and venue for each piece depending on order-specific parameters (e.g., the desired horizon for the execution), algorithm-specific parameters that are estimated from historical data, real-time market data, and feedback about the executions of earlier pieces.

The key characteristic of agency algorithms is that the choice of which stock to trade and how much to buy or sell is made by a portfolio manager who has an investing (rather than trading) horizon in mind. The algorithms are meant to minimize execution costs relative to a specific benchmark (e.g., volume-weighted average price or market price at the time the order arrives at the trading desk), and they are typically developed by sell-side brokers or independent software vendors to serve buy-side clients. Their ultimate goal is to execute a desired position change. Hence they essentially demand liquidity, even though their strategies might utilize nonmarketable limit orders.

Proprietary Algorithms are more diverse and, relative to agency algorithms, more difficult to concisely characterize. Nonetheless, our primary focus is on a new breed of proprietary algorithms that utilizes extremely rapid response to the market environment. These algorithms can be used to provide liquidity or identify a trading interest in the market and use that knowledge to generate profit. Such strategies that utilize extremely low latencies are only deployed by a small set of traders. For example, the CFTC/SEC report on the "flash crash" identifies 17 high-frequency trading firms. Brogaard (2010) reports that NASDAQ identifies 26 firms as being involved in high-frequency trading, but these firms generate most of the order flow in the market and participate in 77% of NASDAQ trades over his sample period.[11]

Because agency and proprietary algorithms differ in their goals, they differ in the specifications of their algorithms and their technology. Agency algorithms are based on historical estimates of price impact and execution probabilities across multiple trading venues and over time, and often require much less real-time input except for tracking the pieces of the orders they execute. For example, volume-weighted average price

[11] The NASDAQ classification excludes proprietary trading desks of large sell-side firms as well as direct access brokers that specialize in providing services to small high-frequency trading firms, and therefore the total number of traders utilizing such low latency may be somewhat larger.

algorithms attempt to distribute executions over time in proportion to the aggregate trading and achieve the average price for the stock. While some agency algorithms offer functionality such as pegging (e.g., tracking the bid or ask side of the market) or discretion (e.g., converting a nonmarketable limit buy order into a marketable order when the ask price decreases), typical agency algorithms do not require millisecond responses to changing market conditions.

We believe that the clock-time periodicity we have identified in Section III.A is driven by agency algorithms. Some algorithms simply check market conditions and execution status every second (or several seconds) and respond to the changes they encounter. Their orders reach the market with a lag that depends on the configurations and locations of their computers, generating the sample distributions of remainders. The similarities between the 2007 and 2008 samples suggest phenomena that are pervasive and do not disappear over time or in different market conditions.

One might conjecture that these patterns cannot be sustainable because sophisticated algorithms will take advantage of them and eliminate them. While there is no doubt that proprietary algorithms respond to such regularities, these responses only serve to accentuate the clock-time periodicities rather than eliminate them. In other words, as long as someone is sending messages in a periodic manner, their actions will provoke strategic responses by others who monitor the market continuously (the low-latency traders) and these responses will tend to amplify the periodicity. Since some proprietary algorithms supply liquidity to agency algorithms, it is conceivable that clustering at certain times helps agency algorithms execute their orders by increasing available liquidity. Furthermore, the clustering of agency algorithms means that the provision of liquidity by one investor to another at those times is higher even without elevated activity of proprietary algorithms. As such, agency algorithms that operate in calendar time would have little incentive to change, making these patterns we identify in the data persist over time.

In contrast to agency algorithms, the hallmark of high-frequency proprietary algorithms is speed: low-latency capabilities. Their need to respond to market events

distinguishes them from agency algorithms. Therefore, these traders invest in co-location and advanced computing technology to create an edge in strategic interactions. The race to interact with the market environment faster and faster not only requires these traders to invest vast resources in technology, but also creates pressure on the various market centers to upgrade their technology and provide solutions tailored to facilitate low-latency activity. While agency algorithms are used in the service of buy-side investing and hence can be justified by the social benefits often attributed to delegated portfolio management (e.g., diversification), the social benefits of high-frequency proprietary trading are more elusive. If we consider electronic market making to be an extension of traditional market making, it provides the service of bridging the intertemporal disaggregation of order flow in continuous markets. Unlike traditional dealers, however, these high-frequency trading firms have no explicit obligations with respect to market presence or market quality, an issue we further discuss in Section VI.

The social benefits of other types of low-latency trading are more difficult to ascertain. One could view them as aiding price discovery by eliminating transient price disturbances, but such an argument in a millisecond environment is tenuous. After all, at such speeds and for such short intervals it is difficult to determine the price component that constitutes a real innovation to the true value of the security as opposed to a transitory influence. The social utility in identifying buy-side interest and trading ahead of it is even more problematic. As such, the prevalence of low-latency algorithms in today's markets invites the question of whether they harm or improve the market quality perceived by long-term investors. We attempt to answer this question in Section IV, but in order to accomplish this we first need to develop a methodology to identify low-latency activity.

### *III.C. Strategic Runs*

The evidence to this point has emphasized message timing. One would ideally like to track low-latency activity in order to decipher its impact on the market. Before turning to

the methodology we use to track the algorithms, it is instructive to present two particular message sets that we believe are typical.

Panel A of Table 2 is an excerpt from the message file for ticker symbol ADCT on October 2, 2007 beginning at 09:51:57.849 and ending at 09:53:09.365 (roughly 72 seconds). Over this period, there were 35 submissions (and 35 cancellations) of orders to buy 100 shares, and 32 submissions (and 32 cancellations) of orders to buy 300 shares. The pricing of the orders caused the bid quote to rapidly oscillate between $20.04 and $20.05. The difference in order sizes and the brief intervals between cancellations and submissions suggest that the traffic is being generated by algorithms responding to each other.

Panel B of Table 2 describes messages (for the same stock on the same day) between 09:57:18.839 and 09:58:36.268 (about 78 seconds). Over this period, orders to sell 100 shares were submitted (and quickly cancelled) 142 times. During much of this period there was no activity except for these messages. As a result of these orders, the ask quote rapidly oscillated between $20.13 and $20.14.

The underlying logic behind each algorithm that generates such strategic runs of messages is difficult to reverse engineer. The interaction in Panel A could be driven by each algorithm's attempt to position a limit order, given the strategy of the other algorithm, so that it would optimally execute against an incoming marketable order. The pattern of submissions and cancellations in Panel B, however, seems more consistent with an attempt to trigger an action on the part of other algorithms and then interact with them. After all, it is clear that an algorithm that repeatedly submits orders and cancels them within 10 ms does not intend to signal anything to human traders (who would not be able to discern such rapid changes in the limit order book). Such algorithms operate in their own space: they are intended to trigger a response from (or respond to) other algorithms. Activity in the limit order book is dominated nowadays by the interaction between automated algorithms, in contrast to a decade ago when human traders still ruled. How, then, do these algorithms affect the environment that the human traders

observe? How is such activity related to market quality measures computed over minutes rather than milliseconds?

To answer these questions, we construct a measure of low-latency activity by identifying "strategic runs," which are linked submissions, cancellations, and executions that are likely to be parts of a dynamic strategy. Since our data do not identify individual traders, our methodology no doubt introduces some noise into the identification of low-latency activity. We nevertheless believe that other attributes of the messages can used to infer linked sequences. In particular, our "strategic runs" (or simply, in this context, "runs") are constructed as follows. Reference numbers supplied with the data unambiguously link an individual limit order with its subsequent cancellation or execution. The point of inference comes in deciding whether a cancellation can be linked to either a subsequent submission of a nonmarketable limit order or a subsequent execution that occurs when the same order is resent to the market priced to be marketable. We impute such a link when the cancellation is followed within one second by a limit order submission or by an execution in the same direction and for the same size. If a limit order is partially executed, and the remainder is cancelled, we look for a subsequent resubmission or execution of the cancelled quantity. In this manner we construct runs forward throughout the day.

Our procedure links roughly 60 percent of the cancellations in the 2007 sample, and 55 percent in the 2008 sample. Although we allow up to a one second delay from cancellation to resubmission, most resubmissions occur much more promptly. The median resubmission delay in our runs is one millisecond. The length of a run can be measured by the number of linked messages. The simplest run would have three messages, a submission of a nonmarketable limit order, its cancellation, and its resubmission as a marketable limit order that executes immediately (i.e., an "active execution"). The shortest run that does not involve an execution is a limit order that was submitted, cancelled, resubmitted, and cancelled or expired at the end of the day. Our sample periods, however, feature many runs of 10 or more linked messages and the

longest run we identify has 93,243 messages. We identify about 57 million runs in the 2007 sample period and 78 million runs in the 2008 sample period.

Panel A of Table 3 presents summary statistics for the runs. We observe that around 80% of the runs have 3 to 9 messages, but the longer runs (10 or more messages) constitute approximately half of the messages that are associated with strategic runs. The proportion of runs that are (at least partially) executed is 33.57% in 2007 and 27.34% in 2008. Interestingly, 22.74% of the 2007 runs (17.77% in 2008) achieve passive execution (when a nonmarketable limit order in the run is hit by an incoming marketable order). This is notable because it can be interpreted as an average fill rate for runs, and stands in contrast to the fill rate for individual limit orders, which is much lower.[12]

About 10.95% (9.64%) of the runs in the 2007 (2008) sample period end with a switch to active execution. That is, a limit order is cancelled and replaced with a marketable order. These numbers attest to the importance of strategies that pursue execution in a gradual fashion. In the combined 2007 and 2008 samples there are a total of 57,848,674 executions. There were (combined) 13,799,814 runs that realized active executions. Since all runs by definition start with a nonmarketable limit order, we can determine that 23.9% (13,799,814/57,848,674) of all executions were preceded by an attempt to obtain a passive execution. This highlights the fluidity with which liquidity suppliers and demanders, often modeled as distinct populations, can in fact switch roles.

Our methodology to impute links between orders no doubt results in misclassifications that introduce an error into the analysis. It is certainly possible that a given cancellation is erroneously linked with a subsequent limit order submitted by a different trader. For a run with many resubmissions to arise solely as an artifact of such errors, however, there would have to be an unbroken chain of spurious linkages. This suggests that longer runs are likely to be more reliable depictions of the activity of actual algorithms than shorter runs. We therefore use runs of 10 or more messages to construct a measure of low-latency traders that we use in the rest of the analysis. While the 10-

[12] The low fill rate of limit orders seems to characterize the modern electronic limit order book environment. Hasbrouck and Saar (2009) report a fill rate of 7.99% for a 2004 sample of Inet data.

message cutoff is somewhat arbitrary, these runs represent about a half of the total number of messages that are linked to runs in each sample period, and we also believe that such longer runs characterize much low-latency activity.

Panel B of Table 3 shows the elapsed time from the beginning to the end of runs of 10 or more messages. It is interesting to note that many of the runs between 10 and 99 messages start and end within a tenth of a second (there are 497,317 such runs in 2007 and 180,675 in 2008), though in general time to completion of a run increases in the number of messages.

## IV. Low-Latency Trading and Market Quality

Agents who engage in low-latency trading and interact with the market over millisecond horizons are at one extreme in the continuum of market participants. Most investors either cannot or choose not to engage the market at this speed.[13] These investors' experience with the market is still best described with the traditional market quality measures in the market microstructure arsenal. Hence, it is natural to ask, how does low-latency activity with its algorithms that interact in milliseconds relate to depth in the market or the range of prices that can be observed over minutes or hours? This question does not have an obvious answer. It seems to resemble the challenge faced by physicists when attempting to relate quantum mechanics' subatomic interactions to our daily life that appears to be governed by Newtonian mechanics. However, if we believe that healthy markets need to attract longer-term investors whose beliefs and preferences are essential for the determination of market prices, then market quality should be measured using time intervals that are easily observed by these investors.

We therefore seek to characterize the influence of low-latency trading on measures of liquidity and short-term volatility observed over 10-minute intervals throughout the day. Measures such as the range between high and low prices in these

[13] The recent SEC Concept Release on Equity Market Structure refers in this context to "long-term investors … who provide capital investment and are willing to accept the risk of ownership in listed companies for an extended period of time" (p. 33).

intervals, the effective and quoted spreads, and the depth of the exchange's limit order book should give us a sense of market quality. And while we would likely not capture every instance of high-frequency proprietary algorithms in each interval of time, the strategic runs we have identified in the previous section could be used to construct a measure of low-latency activity.

### *IV.A. Measures and Methodology*

To measure the intensity of low-latency activity in a stock in each ten-minute interval we use the time-weighted average of the number of strategic runs of 10 messages or more the stock experiences in the interval (*RunsInProcess*).[14] Higher values of *RunsInProcess* indicate greater low-latency activity.

We use our NASDAQ order-level data to compute several measures that represent different aspects of market quality: a measure of short-term volatility and three measures of liquidity. The first measure, *HighLow*, is defined as the highest midquote in an interval minus the lowest midquote in the same interval. The second measure, *EffSprd*, is the average effective spread (or total price impact) of all trades on NASDAQ during the ten-minute interval (where the effective spread of a trade is computed as the absolute value of the difference between the transaction price and the prevailing midquote). The third measure, *Spread*, is the time-weighted average quoted spread (ask price minus the bid price) on the NASDAQ system in an interval. The fourth measure, *NearDepth*, is the time-weighted average number of (visible) shares in the book up to 10 cents from the best posted prices.

Although a ten-minute window is a reasonable interval over which to average the market quality measures, it is sufficiently long (particularly for the low-latency traders) that the analysis must confront the issue of simultaneity. For example, while we aim to

[14] The time-weighting of this measure works as follows. Suppose we construct this variable for the interval 9:50:00am-10:00:00am. If a strategic run started at 9:45:00am and ended at 10:01:00am, it was active for the entire interval and hence it adds 1 to the *RunsInProcess* measure. A run that started at 9:45:00am and ended at 9:51:00am was active for one minute (out of ten) in this interval, and hence adds 0.1 to the measure. Similarly, a run that was active for 6 seconds within this interval adds 0.01.

test whether low-latency trading affects short-term volatility, it is quite possible that short-term volatility attracts or deters low-latency activity and hence affects the number of runs that we can observe in the interval.

To address this problem we propose a two-equation simultaneous equation model in which one of the endogenous variables is *RunsInProcess* (our low-latency activity measure) and the other endogenous variable is the market quality measure (i.e., we estimate the model separately for *HighLow, EffSprd, Spread, and NearDepth*). This variable is indicated in the specifications by the placeholder *MktQuality*. The key to estimating such a model is to identify an instrument for *RunsInProcess* that is not directly affected by market quality in the stock and an instrument for market quality that is not directly affected by the our measures of strategic runs on NASDAQ.

As an instrument for $RunsInProcess_{i,t}$ (the number of runs of 10 messages or more in stock *i* in interval *t*) we use the average number of runs of 10 messages or more in the same interval for the other stocks in our sample (excluding stock *i*), denoted $RunsNotI_t$. This instrument is determined by the number of low-latency trading firms and how active they are in the market during that interval, but it does not utilize information about stock *i* and hence is not directly affected by the liquidity or volatility of stock *i* in interval *t*, rendering it an appropriate instrument. This instrument would be of lower quality if many low-latency proprietary algorithms implemented strategies that require concurrent executions in multiple stocks (i.e., if the liquidity of stock *i* is affected by an algorithm that executes trades in stock *j*). While such algorithms are common in regular quantitative trading (e.g., pairs' trading), our understanding is that they do not dominate trading at extremely low latencies.

As an instrument for market quality we use a measure that is closely related to the liquidity of the stock in the interval, but is not computed from NASDAQ data. Our chief measure is the dollar effective spread (absolute value of the distance between the transaction price and the midquote) computed for the same stock and during the same time interval only from trades executed on other (non-NASDAQ) trading venues. This variable is denoted $EffSprdNotNAS_{i,t}$, and is computed using the TAQ database. It

reflects the general liquidity of the stock in the interval, but it does not utilize information about NASDAQ activity and hence would not be directly determined by the number of strategic runs that are taking place on the NASDAQ system.

This instrument would be of lower quality if many low-latency algorithms pursue cross-market strategies in the same security (i.e., if the same algorithm executes trades on both NASDAQ and another market). A cross-market strategy, however, cannot operate at the lowest latencies because an algorithmic program cannot be co-located at more than one market. This necessarily puts cross-market strategies at a disadvantage relative to co-located single-market algorithms. At least at the lowest latencies, therefore, we believe that the single-market algorithms are dominant.[15] Jovanovic and Menkveld (2010), for example, investigate cross-trading activity in Dutch index stocks between Chi-X and Euronext. In their paper, the time series plot of the cross-trader's imputed net position on a typical day is dominated by components that persist over many minutes, implying operation over periods substantially longer than the low-latency horizons considered here. We elaborate on this issue in Section VI. Considerations of liquidity in multiple markets are also common in agency algorithms that create a montage of the fragmented marketplace to guide their order routing logic. These, however, most likely do not give rise to the long strategic runs that we use to measure the activity of proprietary low-latency traders.

To examine the robustness of our results, we repeat the analysis using another instrument with a similar flavor: the time-weighted average quoted spread from TAQ excluding NASDAQ quotes (denoted $SpreadNotNas_{i,t}$).

With these instruments, we use Two-Stage-Least-Squares (2SLS) to estimate the following two-equation simultaneous equation model for each market quality measure:

$$MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} + e_{1,t}$$
$$RunsInProccess_{i,t} = b_1 MktQuality_{i,t} + b_2 RunsNotI_{i,t} + e_{2,t}$$

[15] Conversations with a NASDAQ official provided support to this view.

where $i = 1, ..., N$ indexes firms, $t = 1, ..., T$ indexes 10-minute time intervals, and *MktQuality* represents one of the market quality measures: *HighLow*, *EffSpread*, *Spread*, and *NearDepth*. All variables are standardized to have a zero mean and unit variance, obviating the need for intercepts in the specification.

The 2SLS methodology effectively replaces $RunsInProcess_{i,t}$ in the first equation with the fitted values from the regression of $RunsInProcess_{i,t}$ on the instruments. Similarly $MktQuality_{i,t}$ in the second equation is replaced with the fitted values of the regression of $MktQuality_{i,t}$ on the instruments. This gives us a consistent estimate of the $a_1$ coefficient that tells us how low-latency activity affects market quality. We estimate the system by pooling observations across all stocks and all time intervals. The standardization of the variables essentially implements a fixed-effects specification. A potential disadvantage of pooling is that the errors of different stocks may not be identically distributed. For robustness, we also report summary measures of the coefficients from stock-by-stock estimations of the system. While stock-by-stock analysis does not assume identically distributed errors across stocks, it leaves us with a much smaller number of observations for each estimation (897 in the 2007 sample period and 819 in the 2008 sample period) and hence has reduced power relative to the pooled time-series/cross-sectional specification.

### *IV.B. Results*

Panel A of Table 4 presents the estimated coefficients of the pooled system side-by-side for the 2007 and 2008 sample periods. The most interesting coefficient is $a_1$, which measures the impact of low-latency activity on the market quality measures. We observe that higher low-latency activity implies lower posted and effective spreads, greater depth, and lower short-term volatility.[16] Moreover, the impact of low-latency activity on market

[16] For robustness, we also carried out all the tests with certain variations on these market quality measures. First, we computed the *EffSpr*d measure only using trades that were not initiated by strategic runs. In other words, we attempted to create a measure of the total price impact that applies to "regular" investors who are not low-latency traders. The results were very similar (in terms of signs and magnitudes of the coefficients as well as their statistical significance) to the measure that includes all NASDAQ trades. Second, we

quality is similar in the 2007 and 2008 sample periods. The fact that low-latency trading decreases short-term volatility and contributes to depth in the 2008 sample period where the market is relentlessly going down and there is heightened uncertainty in the economic environment is particularly noteworthy. It seems to suggest that low-latency activity creates a positive externality in the market at the time that the market needs it the most.

The coefficients on the two instruments have the expected signs and are highly significant. Specifically, the coefficient $a_2$ indicates that when liquidity off NASDAQ is higher, our NASDAQ market quality measures show higher liquidity and lower volatility. Similarly, the coefficient $b_2$ is positive in all specifications, indicating that higher low-latency activity in a specific stock in an interval is associated with higher low-latency activity in other stocks on the NASDAQ system. Finally, the estimated $b_1$ coefficients tell us that low-latency activity is attracted to more liquid and less volatile stocks. Panel B of Table 4 presents roughly similar results from the estimation of the system with $SpreadNotNas_{i,t}$ as the instrument for market liquidity.[17]

While Table 4 presents strong results concerning the impact of low-latency trading on market quality for the entire sample, it could be that this relationship differs for stocks that are somehow fundamentally dissimilar, like small versus large market capitalization stocks. Table 5 presents system estimates in subsamples consisting of four quartiles ranked by the average market capitalization over the sample period.[18] There is not much pattern across the quartiles in the manner low-latency activity affects short-term volatility in the 2007 sample period. The picture in the 2008 sample is different: It appears that during more stressful times, low-latency activity helps reduce volatility in smaller stocks more than it does in larger stocks.

---

computed a depth measure defined as the time-weighted average number of shares in the book up to 50 cents (rather than 10 cents) from the best prices, and the results were also similar.

[17] We have also used $Volume_{i,t}$ as another instrument for market quality, and the results were similar to those presented in Table 4.

[18] The results in the table are presented with $EffSprdNotNAS_{i,t}$ as the instrument for the market quality measures. We obtain similar results (with similar patterns across the quartiles) using $SpreadNotNas_{i,t}$ as the instrument.

Another interesting pattern can be observed in the coefficient $b_1$, which tells us how market quality affects low-latency trading. While low-latency activity increases in market quality for larger stocks in the 2007 sample period, no such relationship is found for smaller stocks, where the coefficient has the opposite sign but is not statistically significant. During the stressful period of June 2008, however, the $b_1$ coefficients suggest a different behavior: Higher liquidity encourages low-latency trading in smaller stocks but not in the top quartile of stocks by market capitalization where we observe the opposite pattern (though the absolute magnitude of the coefficient in large cap stocks is rather small and hence the effect is probably not very strong).

We also estimate the simultaneous equation model separately for subsamples formed as quartiles of NASDAQ's market share of traded volume. Trading in the U.S. occurs on multiple venues, including competing exchanges, crossing networks and Electronic Communications Networks. This fragmentation might jointly affect market quality and low-latency activity. Our results (not reported here), however, show no significant pattern across market-share quartiles. In other words, the beneficial impact of low-latency trading on market quality measures is similar for stocks that have varying degrees of trading concentration on the NASDAQ system.

Lastly, Table 6 shows summary statistics for the stock-by-stock estimations. The results suggest similar conclusions concerning the influence of low-latency trading on market quality. In particular, an increase in low-latency activity decreases short-term volatility, decreases quoted spreads, and increases displayed depth in the limit order book. This is true both in the 2007 and 2008 sample periods. The median coefficient is insignificant when the liquidity measure is *EffSprd* in both sample periods. The only consistent difference between the pooled estimation and the stock-by-stock analysis is that many of the median coefficients of $b_1$ are not statistically significant. In other words, while the impact of low-latency trading on market quality seems robust, our finding that low-latency activity is attracted to more liquid and less volatile stocks should be somewhat qualified due to the insignificant results in the stock-by-stock analysis.

## V. Related Literature

Our paper can be viewed from two related angles: speed of interaction and information dissemination in financial markets, and the characteristics of algorithmic trading and its impact on the market environment. The academic literature in both areas is in its infancy, but there are nonetheless several papers that are related to our study.

Regarding speed, Hendershott and Moulton (2009) look at the introduction of the NYSE's Hybrid Market in 2006, which expanded automatic execution and reduced the execution time for NYSE market orders from ten seconds to under a second. They find that this reduction in latency resulted in worsened liquidity (e.g., spreads increased) but improved informational efficiency. However, Riordan and Storkenmaier (2008) find that a reduction in latency (from 50 to 10 ms) on the Deutsche Boerse' Xetra system is associated with improved liquidity. It could be that the impact of a change in latency on market quality depends on how exactly it affects competition among liquidity suppliers (e.g., the entrance of electronic market makers who can add liquidity but also crowed out traditional liquidity providers) and the level of sophistication of liquidity demanders (e.g., their adoption of algorithms to implement dynamic limit order strategies that can both supply and demand liquidity). Easley, Hendershott, and Ramadorai (2009) examine a change in trading technology on the NYSE in 1980 that increased both the speed and the transparency of the market and find improved liquidity that they attribute to increased competition from off-exchange traders who were better able to compete with the specialists and floor brokers.[19]

[19]Cespa and Foucault (2008) and Easley, O'Hara, and Yang (2010) provide theoretical models in which some traders observe market information with a delay. The two papers employ rather different modeling approaches resulting in somewhat conflicting implications on the impact of differential information latency on the cost of capital, liquidity, and the efficiency of prices. Boulatov and Dierker (2007) investigate information latency from the exchange's perspective: how can the exchange maximize data revenue? Their theoretical model suggests that selling real-time data can be detrimental to liquidity but at the same time enhances the informational efficiency of prices. Moallemi and Sağlam (2010) discuss optimal order placement strategy for a seller facing random exogenous buyer arrivals. In their model, the seller pursues a pegging strategy, and the delayed monitoring caused by latency leads to costly tracking errors.

Algorithmic traders on the Xetra system can attach an order flag that indicates an algorithmic source.[20] Gsell (2008) shows that the majority of orders generated by algorithms demand rather than supply liquidity and are smaller than those sent by human traders, while Groth (2009) finds that algorithmic orders have a higher execution rate than non-algorithmic orders. Gsell and Gomber (2008) show evidence consistent with pegging strategies, and Prix, Loistl, and Huetl (2007), like us, attempt to impute algorithmic strategies. They note that there are certain regularities in the activity of these algorithms, some of which tend to cycle every 60 seconds. Hendershott and Riordan (2009) look at the 30 DAX stocks and find that algorithmic trades have a larger price impact than non-algorithmic trades and seem to contribute more to price discovery.[21]

Three papers focusing on U.S. markets are closely related to our study. Hendershott, Jones, and Menkveld (2009) use the arrival rate of electronic messages on the NYSE as a measure of combined agency and proprietary algorithmic activity. Using an event study approach around the introduction of autoquoting by the NYSE in 2003, the authors find that increase in normalized message count (their proxy for algorithmic trading) impacts liquidity only for large stocks. For these stocks, quoted and effective spreads decline, while quoted depth decreases. The largest stocks also experience improved price discovery. We, on the other hand, find an improvement in market quality using all measures, including depth and short-term volatility, and for all stocks rather than just the largest stocks.[22] Two considerations could account for the difference in

[20] The flag is based on self-reporting, but firms have a fee incentive to identify themselves as algorithmic traders and hence these papers assume that most algorithmic trading is captured by this flag.

[21] There are studies of algorithmic trading outside of U.S. and German equity markets. Chaboud, Chiquoine, Hjalmarsson, and Vega (2009) examine algorithmic trading in the interdealer foreign exchange market. Using an instrument for algorithmic trading measured on a monthly frequency, they find no evidence of a causal relationship between algorithmic trading and increased exchange rate volatility. Jovanovic and Menkveld (2010) provide theoretical and empirical analyses of intermediation in limit order markets. They identify one dealer in Dutch stocks that appears to be implementing automated liquidity provision strategies across two trading platforms, Chi-X and Euronext. They find mixed evidence on the question of whether the activity of the dealer helps or hurts investors.

[22] The average market capitalization (in billion dollars) of sample quintiles reported in Table 1 of Hendershott, Jones, and Menkveld (2009) is 28.99, 4.09, 1.71, 0.90, and 0.41. This corresponds rather well to our sample where the average market capitalization of quintiles is 21.4, 3.8, 2.1, 1.4, and 1.0, though we may have fewer very large and very small stocks compared to their sample.

findings. Firstly, our measure of low-latency trading is designed to capture the activity of high-frequency proprietary algorithms rather than that of agency algorithms. Secondly, prior to the NYSE's introduction of Hybrid Market in 2006, specialists may have faced less competition from high-frequency proprietary akgorithms. The 2003 autoquoting change, therefore, may have mostly affected the activity of agency algorithms.

In a contemporaneous paper, Brogaard (2010) investigates the impact of high-frequency trading on market quality using a dataset containing the activities of 26 high-frequency traders in 120 stocks. He reports that high-frequency traders contribute to liquidity provision in the market, that their trades help price discovery more than trades of other market participants, and that their activity appears to lower volatility. His results, therefore, complement our findings on market quality measures in Section IV, which is especially important given the differences in the design of the experiments in the two papers.

There is no doubt that Brogaard's data on the 26 traders is of high quality: he observes their actual trading activity. On the other hand, his data covers only a subset of firms that utilize low-latency algorithms.[23] Since our measure of low-latency trading relies on imputed strategic runs, we are more likely to capture a broader picture of high-frequency activity.[24] Another important difference between the two papers is that Brogaard's sample spans one week in February 2010 (over which the NASDAQ Composite Index was basically flat), while our 2008 sample provides insights on what happens at times of declining prices and heightened uncertainty. The ability to study low-latency activity during a stressful period for the market is especially important when the conclusion from the analysis of "normal times" is that these traders improve, rather than harm, market quality.

[23] Brogaard's data do not include two types of proprietary traders that utilize low-latency algorithms. First, they lack the proprietary trading desks of larger, integrated firms like Goldman Sachs or JP Morgan. Second, they ignore small firms that use direct access brokers (such as Lime Brokerage or Swift Trade) that specialize in providing services to high-frequency traders.

[24] This is the reason behind our labeling of these traders "low-latency traders" rather than "high-frequency traders." Unlike one or the other terms that are prevalent in the media, our definition is based on an economic idea: Traders who respond to market events.

We note, though, that traders engaged in low-latency activity could impact the market in a negative fashion at times of extreme market stress. The joint CFTC/SEC report regarding the "flash crash" of May 6, 2010, presents a detailed picture of such an event. The report notes that many high-frequency traders scaled down, stopped, or significantly curtailed their trading at some point during this episode. Furthermore, some of the high-frequency traders escalated their aggressive selling during the rapid price decline, removing significant liquidity from the market and hence contributing to the decline. Our study suggests that such behavior is not representative of the manner in which low-latency activity impacts market conditions outside of such extreme episodes.

The potential impact of high-frequency traders on the market environment is analyzed theoretically in two recent papers. Jarrow and Protter (2011) assume downward slopping demand curves and a speed advantage that high-frequency traders have over other traders. This enables the high-frequency traders to create price trends that they exploit to generate profits. Cartea and Penalva (2011) construct a model in the spirit of Grossman and Miller (1988) except that there are high-frequency traders that interject themselves between the liquidity traders and the market makers. In equilibrium, liquidity traders are worse off in the presence of high-frequency traders and the volatility of market prices increases.

Lastly, our paper relates to the analysis of Hasbrouck and Saar (2009) who present evidence consistent with the implementation of dynamic trading strategies by market participants using order-level data from the INET ECN. Hasbrouck and Saar emphasize how technology changed the nature of the market environment. The present paper provides further evidence on attributes of the millisecond environment and the growing importance of algorithmic trading.

## VI. Conclusions

Our paper makes two contributions. First, it describes the millisecond environment in which equity trading currently occurs. The clock-time periodicities and the manner in

which trading responds to market events over millisecond horizons constitute a fundamental change from the manner in which stock markets operated even a few years ago. Second, we study the impact that low-latency activity has on market quality both during normal market conditions and during a period of declining prices and heightened economic uncertainty. Our conclusion is that increased low-latency activity improves traditional yardsticks for market quality such as liquidity and short-term volatility.

The economic issues associated with latency in financial markets are not new, and the private advantage of relative speed was noted well before the advent of our current millisecond environment:

> For some years prior to [the introduction of the telegraph in 1846], William C. Bridges, a stock broker, together with several others, had maintained a unique private 'telegraph' system between Philadelphia and New York. By the ingenious device of establishing stations on high points across New Jersey, on which signals were given by semaphore in the daytime and by light flashes at night, discerned with the aid of telescopes, information on lottery numbers, stock prices, etc., was conveyed in as short a time as ten minutes between the two cities.
> (Barnes, 1911, p. 9)

Nor are low-latency's effects on price dynamics new concerns:

> Some of the mysterious movements in the stock markets of Philadelphia and New York were popularly ascribed to this pioneer financial news bureau.
> (Barnes, ibid)

What is the real economic cost of a delay? It depends on both the risk borne over the delay duration and the effects on participants' strategies. At current latency levels it is difficult to attach much importance to the former. Consider a hypothetical security with a daily log volatility of 0.03 (roughly corresponding, over 250 trading days, to a 47% annual volatility). If the daily volatility is unconditionally distributed evenly over the 6.5 hour trading day, then the volatility over 10 ms is a negligible 0.2 basis points.

The importance of delay for strategic interactions, however, might be much greater. Suppose that the daily volatility is generated by a single randomly-timed announcement that causes the value to change (equiprobably) by ±3%. This 3% can be

captured by a first-mover who observes the announcement and takes a long or short position against others yet unaware, irrespective of whether his absolute time advantage is one minute or one microsecond.

Furthermore, the market itself creates events in the form of imbalances of supply and demand that could be of value to traders who are fast enough to respond to them. There is no doubt that being faster than others entails private advantage, but is it socially beneficial? The first mover in the case of fundamental news imposes costs on other traders, and high adverse selection costs could cause market failure. The fast traders that take advantage of market events could provide valuable liquidity to those seeking immediacy and hence enhance market quality, but could also step ahead of large orders in the book, thereby imposing costs on other liquidity providers (as described in the specialist context by Seppi (1997)).

The early advocates of electronic markets generally envisioned arrangements wherein all traders would enjoy equal access (see Mendelson and Peake (1979), for example). It is therefore particularly striking how much the essential structure of today's electronic markets resembles that of the floor markets they were supposed to have superseded. The old floor-based exchanges (like the NYSE) had a limited number of memberships ("seats"), and only by purchasing or renting a seat could a trader gain access to the floor. Floor-traders had a significant timing advantage over off-floor traders. The modern exchange is essentially a rack-mounted server. The enclosure has a limited number of slots, and only by renting a slot can a trader gain co-located access to the market. Co-located traders have a significant timing advantage over those based elsewhere.

Is this fair? In Regulation Fair Disclosure, the SEC took the stand that firms cannot release fundamental information to a subset of investors before others. On the other hand, Rule 603(a) established a different approach to market data, whereby market

centers could sell data directly to subscribers, in effect creating a tiered system of investors with respect to access to information about market events.[25]

It was also hoped that electronic markets would promote direct interaction of buyers and sellers. There is some evidence, however, that the reverse has occurred. NYSE specialists (the designated market makers on the exchange floor) had a participation rate of 25.3% of the volume just a decade before our sample period.[26] Brogaard (2010) reports that high-frequency traders participate in 73.7% of NASDAQ volume. It is possible that the resulting increase in intermediation is actually desirable in today's fast paced financial markets. If investors do not tolerate delay when trading, it is difficult to assure instantaneous execution without intermediation. And if competition in electronic market making forces the pricing of dealer services to their marginal cost, one argument goes, what is the harm in increased intermediation? One problem with this argument is that the new high-frequency trading firms are not subject to the affirmative and negative obligations that bound formally designated market makers.

In the face of transient supply and demand, NYSE specialists were obligated to stabilize prices and maintain continuous presence in the market. They were subject to restrictions on reaching across the market to take liquidity (i.e., making destabilizing trades). Low-latency traders have no such obligations. Their efficiency and lack of obligations could therefore drive traditional suppliers of liquidity out of business by gaining at their expense in normal times. As a result, at times of severe market stress, low-latency traders can simply step away from the market, causing fragility that did not previously exist.

One of the contributions of our study is the finding that at times of declining prices and heightened economic uncertainty, the nature of the millisecond environment and the positive influence of low-latency activity on market quality remains. However,

[25] Rule 603(a) prohibits an SRO or a broker-dealer from supplying the data via direct feeds faster than it supplies it to the Securities Industry Automation Corporation (SIAC) that processes the data and distributes the "tape." However, the operation of processing and retransmitting data via SIAC appears to add 5 to 10 millisecond and hence subscribers to direct exchange data feeds "see" the information before others who observe the tape.

[26] See New York Stock Exchange Fact Book 1998 Data.

we cannot rule out the possibility of a sudden and severe market condition in which the lack of obligations would result in a market failure. The experience of the "flash crash" in May of 2010 demonstrates that such fragility is certainly possible when a few big players step aside and nobody remains to post limit orders.

Lastly, we believe that it is important to recognize that guaranteeing equal access to market data when the market is both continuous and fragmented (as presently in the U.S.) may be physically impossible. First, Gode and Sunder (2000) claim that when traders are dispersed geographically, transmission delays are sufficiently large to prevent equitable access to a continuous market. Our evidence on the speed of execution against improved quotes suggests that some players are responding within 2-3 ms, while the New York and Chicago round trip (1159 km) is about 8 ms even at the speed of light.

Second, even if one views co-location as the ultimate equalizer of dispersed traders, it leads to the impossibility of achieving equal access in fragmented markets. Since the same stock is traded on multiple trading venues, a co-located computer near the servers of exchange A would be at a disadvantage in responding to market events in the same securities on exchange B compared to computers co-located with exchange B. Hence, unless markets change from continuous to periodic, some traders will always have lower latency than others. Our findings in this paper suggest that in the current (post Reg NMS) environment, increased low-latency activity need not invariably work to the detriment of long-term investors.

**References**


Barnes, A. W., 1911, *History of the Philadelphia Stock Exchange*, Philadelphia, Cornelius Baker.

Bergan, Peter, and Colleen Devine, 2005, Algorithmic trading: What should you be doing? In *Algorithmic Trading: Precision, Control, Execution* (Brian R. Bruce, editor), Institutional Investor.

Boulatov, Alex, and Martin Dierker, 2007, Pricing prices, Working paper, University of Houston.

Brogaard, Jonathan A., 2010, High frequency trading and its impact on market quality, Working paper, Northwestern University.

Cartea, Alvaro, and Jose Penalva, 2011, Where is the value in high frequency trading? Working paper, Universidad Carlos III de Madrid.

Cespa, Giovanni, and Thierry Foucault, 2008, Insiders-outsiders, transparency, and the value of the ticker, Working paper, Queen Mary University of London and HEC.

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

Donefer, Bernard S., 2010, Algos gone wild: Risk in the world of automated trading strategies, *Journal of Trading* 5 (2), 31-34.

Easley, David, Terrence Hendershott, and Tarun Ramadorai, 2009, Leveling the trading field, Working paper, Cornell University.

Easley, David, Maureen O'Hara, and Liyan Yang, 2010, Differential access to price information in financial markets, Working paper, Cornell University.

Engle, Robert F., and Jeffrey R. Russell, 1998, Autoregressive conditional duration: A new model for irregularly spaced transaction data, *Econometrica* 66, 1127-1162.

Gode, Dhananjay K. and Shyam Sunder, 2000, On the impossibility of equitable continuously-clearing markets with geographically distributed traders, Working paper, New York University.

Groth, Sven S., 2009, Further evidence on "Technology and liquidity provision: The blurring of Tradition Definitions," Working paper, Goethe University, Frankfurt am Main.

Gsell, Markus, 2009, Algorithmic activity on Xetra, *Journal of Trading* 4, 74-86

Gsell, Markus, and Peter Gomber, 2008, Algorithmic trading versus human traders—Do they behave different in securities markets? Working paper, Goethe University, Frankfurt am Main.

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

Hautsch, Nikolaus, 2004, *Modeling Irregularly Spaced Financial Data: Theory and Practice of Dynamic Duration Models*, Springer.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2009, Does algorithmic trading improve liquidity? *Journal of Finance*, forthcoming.

Hendershott, Terrence, and Pamela C. Moulton, 2009, Speed and stock market quality: The NYSE's Hybrid, Working paper, University of California at Berkeley.

Hendershott, Terrence, and Ryan Riordan, 2009, Algorithmic trading and information, Working paper, University of California at Berkeley.

Jarrow, Robert A., and Philip Protter, 2011, A dysfunctional role of high frequency trading in electronic markets, Working paper, Cornell University.

Jovanovic, Boyan, and Albert J. Menkveld, 2010, Middlemen in limit-order markets, Working paper, New York University.

Kosinski, R. J., 2010, A literature review on reaction time, Working paper, Clemson University.

Lauricella, Tom, and Jenny Strasburg, 2010, SEC probes cancelled trades, *The Wall Street Journal*, September 1, A1.

Mendelson, Morris, and Junius W. Peake, 1979, The ABCs of trading on a national market system, *Financial Analysts Journal* 35, 31-34+27-42.

Moallemi, Ciamac C., and Mehmet Sağlam, 2010, The cost of latency, Working paper, Columbia University.

O'Hara, Maureen, and Mao Ye, 2010, Is market fragmentation harming market quality? *Journal of Financial Economics*, forthcoming.

Prix, Johannes, Otto Loistl, and Michael Huetl, 2007, Algorithmic trading patterns in Xetra orders, *European Journal of Finance* 13, 717-739.

Riordan, Ryan, and Andreas Storkenmaier, 2008, Optical illusions: The effects of exchange system latency on liquidity, Working paper, University of Karlsruhe, Germany.

Securities and Exchange Commission, 2010, *Concept Release on Equity Market Structure* (Release No. 34-61358).

Seppi, Duane J. (1997). Liquidity provision with limit orders and a strategic specialist. *Review of Financial Studies* 10(1): 103-150

U.S. Commodities Futures Trading Commission, and U.S. Securities and Exchange Commission, 2010, Findings regarding the market events of May 6, 2010, (Washington D.C.).

# Table 1

## Summary Statistics

This table presents summary statistics for the stocks in our sample. The universe of stocks used in the study is comprised of the 500 largest stocks by market capitalization on September 28, 2007. We investigate trading in these stocks in two sample periods: (i) October 2007 (23 trading days), and (ii) June 2008 (21 trading days). Since the main econometric analysis in the paper requires sufficient level of activity in the stocks, we apply the following screen to the stocks in each sample period: A firm is rejected if the proportion of 10-minute intervals with fewer than 250 messages is above 10%. A "message" for the purpose of this screen could be a submission, a cancellation, or an execution of a limit order. After applying the screen (and dropping Google and Apple due to computational limitations), our sample consists of 345 stocks in the October 2007 sample period and 394 stocks in the June 2008 sample period. In Panel A we report summary statistics from the CRSP database. *MktCap* is the market capitalization of the firms computed using closing prices on the last trading day prior to the start of the sample period. *ClsPrice* is the average closing price, *AvgVol* is the average daily share volume, and *AvgRet* is the average daily return. These variables are averaged across time for each firm, and the table entries refer to the sample distribution of these firm-averages. Panel B presents summary statistics from the NASDAQ market computed using TotalView-ITCH data. We report the average daily number of orders submitted, cancelled, and executed in each sample period, along with the average daily number of shares executed. The summary measures for the limit order book include the time-weighted average depth in the book, the time-weighted average depth near current market prices (i.e., within 10 cents of the best bid or ask prices), and the time-weighted average dollar quoted spread (the distance between the bid and ask prices). We also report the effective (half) spread, defined as the absolute value of the difference between the transaction price and the quote midpoint, averaged across all transactions.

Panel A: CRSP Summary Statistics

| | 2007 | | | | 2008 | | | |
|---|---|---|---|---|---|---|---|---|
| | *MktCap ($Million)* | *ClsPrice ($)* | *AvgVol (1,000s)* | *AvgRet (%)* | *MktCap ($Million)* | *ClsPrice ($)* | *AvgVol (1,000s)* | *AvgRet (%)* |
| Mean | 5,936 | 34.98 | 3,092 | 0.110 | 4,908 | 30.09 | 2,871 | -0.565 |
| Median | 2,069 | 29.07 | 1,074 | 0.123 | 1,648 | 24.67 | 1,116 | -0.512 |
| Std | 18,402 | 25.55 | 7,950 | 0.557 | 16,337 | 27.84 | 6,263 | 0.618 |
| Min | 789 | 2.22 | 202 | -2.675 | 286 | 2.32 | 112 | -3.449 |
| Max | 275,598 | 272.07 | 77,151 | 1.933 | 263,752 | 278.66 | 74,514 | 0.817 |

Panel B. NASDAQ (TotalView-ITCH) Summary Statistics

| | | Number of Limit Order Submissions | Number of Limit Order Cancellations | Number of Marketable Order Executions | Shares Executed (1,000s) | Depth (1,000s) | Near Depth (1,000s) | Quoted Spread ($) | Eff. Half Spread ($) |
|---|---|---|---|---|---|---|---|---|---|
| 2007 | Mean | 41,477 | 37,126 | 3,593 | 1,363 | 243 | 29 | 0.033 | 0.013 |
| | Median | 27,130 | 24,374 | 2,489 | 548 | 74 | 6 | 0.025 | 0.010 |
| | Std | 44,334 | 40,039 | 3,290 | 3,154 | 813 | 129 | 0.031 | 0.011 |
| | Min | 9,658 | 8,013 | 695 | 130 | 13 | 1 | 0.010 | 0.005 |
| | Max | 305,688 | 308,178 | 22,644 | 32,305 | 7,979 | 1,555 | 0.313 | 0.111 |
| 2008 | Mean | 52,756 | 48,671 | 3,546 | 1,177 | 254 | 22 | 0.034 | 0.012 |
| | Median | 34,875 | 31,712 | 2,329 | 486 | 78 | 5 | 0.023 | 0.008 |
| | Std | 54,978 | 50,882 | 3,666 | 2,556 | 886 | 77 | 0.039 | 0.012 |
| | Min | 8,889 | 7,983 | 291 | 42 | 10 | 0 | 0.010 | 0.004 |
| | Max | 401,140 | 409,803 | 28,105 | 32,406 | 12,502 | 1,241 | 0.462 | 0.132 |

# Table 2

## Examples of Strategic Runs for Ticker Symbol ADCT on October 2, 2007

This table presents examples of "strategic runs," which are linked submissions, cancellations, and executions that are likely to be parts of a dynamic strategy of a trading algorithm. The examples are taken from activity in one stock (ATC Telecommunications, ticker symbol ADCT) on October 2, 2007. We identify the existence of these strategic runs by imputing links between different submissions, cancellations, and executions based on direction, size, and timing. In the two cases presented below, the activity in the table constitutes all messages in this stock (i.e., there are no intervening messages that are unrelated to these strategic runs). In Panel A, we present order activity starting around 9:51:57am where two algorithms "play" with each other (i.e., they submit and cancel messages in response to one another). The messages sent by the second algorithm are highlighted in the table. The algorithms are active for one minute and 12 seconds, sending 137 messages (submissions and cancellations) to the market. In Panel B we present order activity starting around 9:57:18am where one algorithm submits and cancels orders. The algorithm is active for one minute and eighteen seconds, sending 142 messages (submissions and cancellations) to the market.

Panel A: ADCT Order Activity Starting 09:51:57.849

| Time | Message | B/S | Shares | Price | Bid | Offer |
|---|---|---|---|---|---|---|
| 09:51:57.849 | Submission | Buy | 100 | 20.00 | 20.03 | 20.05 |
| 09:52:13.860 | Submission | Buy | 300 | 20.03 | 20.03 | 20.04 |
| 09:52:16.580 | Cancellation | Buy | 300 | 20.03 | 20.03 | 20.04 |
| 09:52:16.581 | Submission | Buy | 300 | 20.03 | 20.03 | 20.04 |
| 09:52:23.245 | Cancellation | Buy | 100 | 20.00 | 20.04 | 20.05 |
| 09:52:23.245 | Submission | Buy | 100 | 20.04 | 20.04 | 20.05 |
| 09:52:23.356 | Cancellation | Buy | 300 | 20.03 | 20.04 | 20.05 |
| 09:52:23.357 | Submission | Buy | 300 | 20.04 | 20.04 | 20.05 |
| 09:52:26.307 | Cancellation | Buy | 300 | 20.04 | 20.05 | 20.07 |
| 09:52:26.308 | Submission | Buy | 300 | 20.05 | 20.05 | 20.07 |
| 09:52:29.401 | Cancellation | Buy | 300 | 20.05 | 20.04 | 20.07 |
| 09:52:29.402 | Submission | Buy | 300 | 20.04 | 20.04 | 20.07 |
| 09:52:29.402 | Cancellation | Buy | 100 | 20.04 | 20.04 | 20.07 |
| 09:52:29.403 | Submission | Buy | 100 | 20.00 | 20.04 | 20.07 |
| 09:52:32.665 | Cancellation | Buy | 100 | 20.00 | 20.04 | 20.07 |
| 09:52:32.665 | Submission | Buy | 100 | 20.05 | 20.05 | 20.07 |
| 09:52:32.672 | Cancellation | Buy | 100 | 20.05 | 20.04 | 20.07 |
| 09:52:32.678 | Submission | Buy | 100 | 20.05 | 20.05 | 20.07 |
| 09:52:32.707 | Cancellation | Buy | 100 | 20.05 | 20.04 | 20.07 |
| 09:52:32.708 | Submission | Buy | 100 | 20.05 | 20.05 | 20.07 |
| 09:52:32.717 | Cancellation | Buy | 100 | 20.05 | 20.04 | 20.07 |
| 09:52:32.745 | Cancellation | Buy | 300 | 20.04 | 20.04 | 20.07 |
| 09:52:32.745 | Submission | Buy | 100 | 20.05 | 20.05 | 20.07 |
| 09:52:32.746 | Submission | Buy | 300 | 20.05 | 20.05 | 20.07 |
| 09:52:32.747 | Cancellation | Buy | 100 | 20.05 | 20.05 | 20.07 |
| 09:52:32.772 | Submission | Buy | 100 | 20.02 | 20.05 | 20.07 |
| 09:52:32.776 | Cancellation | Buy | 300 | 20.05 | 20.04 | 20.07 |
| 09:52:32.777 | Cancellation | Buy | 100 | 20.02 | 20.04 | 20.07 |
| 09:52:32.777 | Submission | Buy | 300 | 20.04 | 20.04 | 20.07 |
| 09:52:32.778 | Submission | Buy | 100 | 20.05 | 20.05 | 20.07 |
| 09:52:32.778 | Cancellation | Buy | 300 | 20.04 | 20.05 | 20.07 |
| 09:52:32.779 | Submission | Buy | 300 | 20.05 | 20.05 | 20.07 |
| 09:52:32.779 | Cancellation | Buy | 100 | 20.05 | 20.05 | 20.07 |
| 09:52:32.807 | Cancellation | Buy | 300 | 20.05 | 20.04 | 20.07 |
| 09:52:32.808 | Submission | Buy | 100 | 20.02 | 20.04 | 20.07 |
| 09:52:32.808 | Submission | Buy | 300 | 20.04 | 20.04 | 20.07 |
| 09:52:32.809 | Cancellation | Buy | 100 | 20.02 | 20.04 | 20.07 |

… the interaction between the two strategic runs continues for 95 additional messages until a limit order of the 300-share run is executed by an incoming marketable order at 09:53:09.365.

Panel B: ADCT Order Activity Starting 09:57:18.839

| Time | Message | B/S | Shares | Price | Bid | Ask |
|---|---|---|---|---|---|---|
| 09:57:18.839 | Submission | Sell | 100 | 20.18 | 20.11 | 20.14 |
| 09:57:18.869 | Cancellation | Sell | 100 | 20.18 | 20.11 | 20.14 |
| 09:57:18.871 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:18.881 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:18.892 | Submission | Sell | 100 | 20.16 | 20.11 | 20.14 |
| 09:57:18.899 | Cancellation | Sell | 100 | 20.16 | 20.11 | 20.14 |
| 09:57:18.902 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:18.911 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:18.922 | Submission | Sell | 100 | 20.16 | 20.11 | 20.14 |
| 09:57:18.925 | Cancellation | Sell | 100 | 20.16 | 20.11 | 20.14 |
| 09:57:18.942 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:18.954 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:18.958 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:18.961 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:18.973 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:18.984 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:18.985 | Submission | Sell | 100 | 20.16 | 20.11 | 20.14 |
| 09:57:18.995 | Cancellation | Sell | 100 | 20.16 | 20.11 | 20.14 |
| 09:57:18.996 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:19.002 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:19.004 | Submission | Sell | 100 | 20.16 | 20.11 | 20.14 |
| 09:57:19.807 | Cancellation | Sell | 100 | 20.16 | 20.11 | 20.13 |
| 09:57:19.807 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.451 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.461 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.471 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.480 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.481 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.484 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.499 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.513 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.521 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.532 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.533 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.542 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.554 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.562 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.571 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.581 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.592 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.601 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.611 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.622 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.667 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.671 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.681 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.742 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |
| 09:57:20.756 | Cancellation | Sell | 100 | 20.13 | 20.11 | 20.14 |
| 09:57:20.761 | Submission | Sell | 100 | 20.13 | 20.11 | 20.13 |

… the strategic run continues for 89 additional messages until it stops at 09:58:36.268.

# Table 3

## Strategic Runs

This table presents summary statistics for "strategic runs," which are linked submissions, cancellations, and executions that are likely to be parts of a dynamic strategy. The imputed links between different submissions, cancellations, and executions are based on direction, size, and timing. Specifically, when a cancellation is followed within one second by a submission of a limit order in the same direction and for the same quantity, or by an execution in the same direction and for the same quantity, we impute a link between the messages. The methodology that tracks the strategic runs also takes note of partial executions and partial cancellations of orders. In Panel A we sort runs into categories by length (i.e., the number of linked messages), and report information about the number of runs, messages, and executions (separately active and passive) within each category. An active execution is when the run ends with a marketable limit order that executes immediately. A passive execution is when a standing limit order that is part of a run is executed by an incoming marketable order. One run could potentially result in both a passive execution and an active execution if the passive execution did not exhaust the order, and the reminder was cancelled and resubmitted to generate an immediate active execution. Panel B shows the elapsed time from the beginning to the end of runs of 10 or more messages, which are the runs that we use to construct our measure of low-latency activity.

Panel A: Summary Statistics of Strategic Runs

| | Length Of Runs | Runs (#) | Runs (%) | Messages (#) | Messages (%) | Active Exec. (#) | Active Exec. Rate | Passive Exec. (#) | Passive Exec. Rate | Total Exec. (#) | Total Exec. Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2007 | 3-4 | 27,344,930 | 47.99% | 105,690,858 | 22.53% | 3,720,292 | 13.61% | 5,476,480 | 20.03% | 9,172,711 | 33.54% |
| | 5-9 | 17,998,854 | 31.59% | 118,037,347 | 25.17% | 1,882,712 | 10.46% | 4,941,592 | 27.46% | 6,798,313 | 37.77% |
| | 10-14 | 6,560,499 | 11.51% | 75,353,085 | 16.07% | 284,960 | 4.34% | 1,468,072 | 22.38% | 1,744,893 | 26.60% |
| | 15-19 | 1,842,320 | 3.23% | 30,948,629 | 6.60% | 173,262 | 9.40% | 418,977 | 22.74% | 589,789 | 32.01% |
| | 20-99 | 3,073,546 | 5.39% | 100,494,251 | 21.43% | 172,094 | 5.60% | 619,304 | 20.15% | 787,245 | 25.61% |
| | 100+ | 160,903 | 0.28% | 38,503,154 | 8.21% | 6,529 | 4.06% | 31,316 | 19.46% | 37,508 | 23.31% |
| | All | 56,981,052 | 100.00% | 469,027,324 | 100.00% | 6,239,849 | 10.95% | 12,955,71 | 22.74% | 19,130,459 | 33.57% |
| 2008 | 3-4 | 40,284,620 | 51.35% | 156,714,747 | 26.25% | 4,459,563 | 11.07% | 5,916,127 | 14.69% | 10,355,650 | 25.71% |
| | 5-9 | 23,744,638 | 30.27% | 155,608,785 | 26.06% | 2,297,553 | 9.68% | 5,324,835 | 22.43% | 7,599,729 | 32.01% |
| | 10-14 | 8,262,256 | 10.53% | 94,723,010 | 15.87% | 354,704 | 4.29% | 1,600,453 | 19.37% | 1,948,080 | 23.58% |
| | 15-19 | 2,295,030 | 2.93% | 38,561,692 | 6.46% | 221,307 | 9.64% | 451,793 | 19.69% | 671,084 | 29.24% |
| | 20-99 | 3,696,434 | 4.71% | 118,816,877 | 19.90% | 219,686 | 5.94% | 627,419 | 16.97% | 844,207 | 22.84% |
| | 100+ | 160,661 | 0.20% | 32,615,369 | 5.46% | 7,152 | 4.45% | 22,687 | 14.12% | 29,695 | 18.48% |
| | All | 78,443,639 | 100.00% | 597,040,480 | 100.00% | 7,559,965 | 9.64% | 13,943,314 | 17.77% | 21,448,445 | 27.34% |

Panel B: Distribution of Elapsed Time for Runs of 10 or more Messages

| | | | Elapsed Time | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Length of Run | Number of Runs | < 0.1 sec. | [0.1,1) sec. | [1,60) sec. | [1,10) min. | [10,60) min. | > 60 min. |
| 2007 | 10-14 | 6,560,499 | 276,703 | 353,093 | 3,015,701 | 2,386,218 | 462,458 | 66,326 |
| | 15-19 | 1,842,320 | 73,978 | 93,759 | 763,002 | 716,794 | 172,526 | 22,261 |
| | 20-99 | 3,073,546 | 124,008 | 218,861 | 1,075,282 | 1,109,339 | 458,586 | 87,470 |
| | 100-999 | 158,032 | 218 | 16,827 | 43,277 | 32,977 | 24,090 | 40,643 |
| | 1,000-4,999 | 2,523 | 0 | 0 | 1,392 | 609 | 263 | 259 |
| | 5,000+ | 348 | 0 | 0 | 126 | 134 | 30 | 58 |
| | All | 11,637,268 | 474,907 | 682,540 | 4,898,780 | 4,246,071 | 1,117,953 | 217,017 |
| 2008 | 10-14 | 8,262,256 | 109,077 | 164,355 | 3,785,673 | 3,572,232 | 560,216 | 70,703 |
| | 15-19 | 2,295,030 | 25,984 | 34,601 | 842,787 | 1,148,372 | 218,637 | 24,649 |
| | 20-99 | 3,696,434 | 38,955 | 74,953 | 987,683 | 1,791,617 | 694,245 | 108,981 |
| | 100-999 | 159,401 | 45 | 5,613 | 32,396 | 35,553 | 32,696 | 53,098 |
| | 1,000-4,999 | 1,211 | 0 | 0 | 600 | 442 | 83 | 86 |
| | 5,000+ | 49 | 0 | 0 | 16 | 21 | 5 | 7 |
| | All | 14,414,381 | 174,061 | 279,522 | 5,649,155 | 6,548,237 | 1,505,882 | 257,524 |

# Table 4

## Simultaneous Equation Model: Low-Latency Trading and Market Quality

This table presents analysis of the manner in which low-latency trading affects market quality. To measure the intensity of low-latency activity in a stock in each ten-minute interval, we use the time-weighted average of the number of strategic runs of 10 messages or more the stock experiences in the interval (*RunsInProcess*). We use NASDAQ order-level data to compute several measures that represent different aspects of market quality on the NASDAQ system in each time interval: (i) *HighLow* is the highest midquote minus the lowest midquote in the same interval, (ii) *EffSprd* is the average effective spread (or total price impact) of a trade, computed as the absolute value of the difference between the transaction price and the prevailing midquote, (iii) *Spread* is the time-weighted average quoted spread (ask price minus the bid price), and (iv) *NearDepth* is the time-weighted average number of (visible) shares in the book up to 10 cents from the best posted prices. Due to the potential simultaneity between market quality and low-latency trading, we estimate the following two-equation simultaneous equation model for *RunsInProcess* and each of the market quality measures (*HighLow, EffSprd, Spread, and NearDepth*):

$$MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} + e_{1,t}$$

$$RunsInProccess_{i,t} = b_1 MktQuality_{i,t} + b_2 RunsNotI_{i,t} + e_{2,t}$$

As an instrument for *RunsInProcess* we use *RunsNotI*, which is the average number of runs of 10 messages or more in the same interval for the other stocks in our sample (excluding stock *i*). In Panel A, we present the results with our main instrument for the market quality measures: *EffSprdNotNas*, which is the average dollar effective spread computed from trades executed in the same stock and during the same time interval on other trading venues (from the TAQ database). For robustness, we present in Panel B the analysis with an alternative instrument, *SpreadNotNas*, which is the time-weighted average quoted spread (from TAQ) that excludes NASDAQ quotes. We estimate the simultaneous equation model by pooling observations across all stocks and all time intervals. To enable a meaningful pooling of data, we standardize each variable by subtracting from each observation the stock-specific time-series average and dividing by the stock-specific time-series standard deviation. Hence, this formulation essentially implements a fixed-effects specification. We report the coefficients and the p-values (against a two-sided alternative) side-by-side for the 2007 and 2008 sample periods.

Panel A: Estimates of the Simultaneous Equation Model with Instruments *EffSprdNotNAS* and *RunsNotI*

| | | 2007 | | | | 2008 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $a_1$ | $a_2$ | $b_1$ | $b_2$ | $a_1$ | $a_2$ | $b_1$ | $b_2$ |
| *HighLow* | Coef. | -0.339 | 0.474 | -0.054 | 0.534 | -0.451 | 0.463 | -0.121 | 0.485 |
| | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| *Spread* | Coef. | -0.501 | 0.572 | -0.044 | 0.532 | -0.531 | 0.551 | -0.101 | 0.485 |
| | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| *EffSprd* | Coef. | -0.179 | 0.396 | -0.065 | 0.537 | -0.121 | 0.233 | -0.245 | 0.497 |
| | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| *NearDepth* | Coef. | 0.444 | -0.217 | 0.114 | 0.516 | 0.644 | -0.138 | 0.334 | 0.402 |
| | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |

Panel B: Estimates of the Simultaneous Equation Model with Instruments *SpreadNotNAS* and *RunsNotI*

| | | 2007 | | | | 2008 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $a_1$ | $a_2$ | $b_1$ | $b_2$ | $a_1$ | $a_2$ | $b_1$ | $b_2$ |
| *HighLow* | Coef. | -0.362 | 0.366 | -0.157 | 0.494 | -0.416 | 0.404 | -0.169 | 0.463 |
| | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| *Spread* | Coef. | -0.254 | 0.744 | -0.080 | 0.513 | -0.177 | 0.797 | -0.090 | 0.490 |
| | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| *EffSprd* | Coef. | -0.177 | 0.330 | -0.179 | 0.507 | -0.082 | 0.225 | -0.316 | 0.486 |
| | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (0.671) | (<.001) | (<.001) | (<.001) |
| *NearDepth* | Coef. | 0.344 | -0.289 | 0.197 | 0.488 | 0.565 | -0.190 | 0.317 | 0.409 |
| | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |

# Table 5

## Low-Latency Trading and Market Quality by Size Quartiles

This table presents the results of a simultaneous equation model of low-latency trading and market quality separately for stocks in each firm-size quartile. To measure the intensity of low-latency activity in a stock in each ten-minute interval, we use the time-weighted average of the number of strategic runs of 10 messages or more the stock experiences in the interval (*RunsInProcess*). We use NASDAQ order-level data to compute several measures that represent different aspects of market quality on the NASDAQ system in each time interval: (i) *HighLow* is the highest midquote minus the lowest midquote in the same interval, (ii) *EffSprd* is the average effective spread (or total price impact) of a trade, computed as the absolute value of the difference between the transaction price and the prevailing midquote, (iii) *Spread* is the time-weighted average quoted spread (ask price minus the bid price), and (iv) *NearDepth* is the time-weighted average number of (visible) shares in the book up to 10 cents from the best posted prices. Due to the potential simultaneity between market quality and low-latency trading, we estimate the following two-equation simultaneous equation model for *RunsInProcess* and each of the market quality measures (*HighLow, EffSprd, Spread, and NearDepth*):

$$MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} + e_{1,t}$$

$$RunsInProccess_{i,t} = b_1 MktQuality_{i,t} + b_2 RunsNotI_{i,t} + e_{2,t}$$

As an instrument for *RunsInProcess* we use *RunsNotI*, which is the average number of runs of 10 messages or more in the same interval for the other stocks in our sample (excluding stock *i*). As an instrument for the market quality measures we use *EffSprdNotNas*, which is the average dollar effective spread computed from trades executed in the same stock and during the same time interval on other trading venues (from the TAQ database). We estimate the simultaneous equation model by pooling observations across all stocks and all time intervals. To enable a meaningful pooling of data, we standardize each variable by subtracting from each observation the stock-specific time-series average and dividing by the stock-specific time-series standard deviation. Hence, this formulation essentially implements a fixed-effects specification. We report the coefficients and the p-values (against a two-sided alternative) side-by-side for the 2007 and 2008 sample periods.

| Dep. Var. | | | 2007 $a_1$ | 2007 $a_2$ | 2007 $b_1$ | 2007 $b_2$ | 2008 $a_1$ | 2008 $a_2$ | 2008 $b_1$ | 2008 $b_2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| *HighLow* | Q1 (small) | Coef. | -0.348 | 0.451 | 0.016 | 0.531 | -0.654 | 0.415 | -0.197 | 0.338 |
| | | p-value | (<.001) | (<.001) | (0.090) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q2 | Coef. | -0.377 | 0.455 | 0.003 | 0.534 | -0.646 | 0.407 | -0.191 | 0.336 |
| | | p-value | (<.001) | (<.001) | (0.712) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q3 | Coef. | -0.334 | 0.475 | -0.033 | 0.533 | -0.455 | 0.464 | -0.127 | 0.484 |
| | | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q4 (large) | Coef. | -0.312 | 0.500 | -0.133 | 0.539 | -0.279 | 0.521 | 0.017 | 0.713 |
| | | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| *Spread* | Q1 (small) | Coef. | -0.562 | 0.569 | 0.013 | 0.532 | -0.742 | 0.486 | -0.169 | 0.339 |
| | | p-value | (<.001) | (<.001) | (0.090) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q2 | Coef. | -0.530 | 0.577 | 0.002 | 0.534 | -0.758 | 0.494 | -0.158 | 0.337 |
| | | p-value | (<.001) | (<.001) | (0.712) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q3 | Coef. | -0.523 | 0.586 | -0.027 | 0.532 | -0.542 | 0.547 | -0.108 | 0.484 |
| | | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q4 (large) | Coef. | -0.437 | 0.562 | -0.117 | 0.534 | -0.334 | 0.625 | 0.014 | 0.713 |
| | | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| *EffSprd* | Q1 (small) | Coef. | -0.185 | 0.357 | 0.020 | 0.530 | -0.140 | 0.166 | -0.524 | 0.360 |
| | | p-value | (<.001) | (<.001) | (0.091) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q2 | Coef. | -0.158 | 0.407 | 0.003 | 0.533 | -0.150 | 0.181 | -0.456 | 0.357 |
| | | p-value | (<.001) | (<.001) | (0.743) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q3 | Coef. | -0.176 | 0.428 | -0.037 | 0.536 | -0.121 | 0.248 | -0.244 | 0.499 |
| | | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q4 (large) | Coef. | -0.193 | 0.379 | -0.177 | 0.543 | -0.092 | 0.306 | 0.029 | 0.712 |
| | | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| *NearDepth* | Q1 (small) | Coef. | 0.423 | -0.188 | -0.039 | 0.537 | 0.769 | -0.088 | 0.584 | 0.214 |
| | | p-value | (<.001) | (<.001) | (0.093) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q2 | Coef. | 0.527 | -0.192 | -0.007 | 0.535 | 0.764 | -0.094 | 0.549 | 0.222 |
| | | p-value | (<.001) | (<.001) | (0.712) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q3 | Coef. | 0.432 | -0.209 | 0.073 | 0.522 | 0.646 | -0.122 | 0.385 | 0.386 |
| | | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |
| | Q4 (large) | Coef. | 0.406 | -0.259 | 0.242 | 0.507 | 0.534 | -0.215 | -0.042 | 0.726 |
| | | p-value | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) | (<.001) |

# Table 6

## Stock-by-Stock Estimation of Simultaneous Equation Model

This table presents the median coefficient estimate (and its p-value) from a stock-by-stock estimation of a simultaneous equation model that we use to examine the manner in which low-latency trading affects market quality. To measure the intensity of low-latency activity in a stock in each ten-minute interval, we use the time-weighted average of the number of strategic runs of 10 messages or more the stock experiences in the interval (*RunsInProcess*). We use NASDAQ order-level data to compute several measures that represent different aspects of market quality on the NASDAQ system in each time interval: (i) *HighLow* is the highest midquote minus the lowest midquote in the same interval, (ii) *EffSprd* is the average effective spread (or total price impact) of a trade, computed as the absolute value of the difference between the transaction price and the prevailing midquote, (iii) *Spread* is the time-weighted average quoted spread (ask price minus the bid price), and (iv) *NearDepth* is the time-weighted average number of (visible) shares in the book up to 10 cents from the best posted prices. Due to the potential simultaneity between market quality and low-latency trading, we estimate the following two-equation simultaneous equation model for *RunsInProcess* and each of the market quality measures (*HighLow, EffSprd, Spread, and NearDepth*):

$$MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} + e_{1,t}$$

$$RunsInProccess_{i,t} = b_1 MktQuality_{i,t} + b_2 RunsNotI_{i,t} + e_{2,t}$$

As an instrument for *RunsInProcess* we use *RunsNotI*, which is the average number of runs of 10 messages or more in the same interval for the other stocks in our sample (excluding stock *i*). In Panel A, we present the results with our main instrument for the market quality measures: *EffSprdNotNas*, which is the average dollar effective spread computed from trades executed in the same stock and during the same time interval on other trading venues (from the TAQ database). For robustness, we present in Panel B the analysis with an alternative instrument, *SpreadNotNas*, which is the time-weighted average quoted spread (from TAQ) that excludes NASDAQ quotes. We standardize each variable by subtracting from each observation the stock-specific time-series average and dividing by the stock-specific time-series standard deviation. Hence, this formulation essentially implements a fixed-effects specification. We estimate the simultaneous equation model for each stock separately, and report the median coefficient (across the stocks) and its p-value.

Panel A: Cross-Sectional Median Coefficient Estimate when Instruments are *EffSprdNotNAS* and *RunsNotI*

| | | 2007 | | | | 2008 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $a_1$ | $a_2$ | $b_1$ | $b_2$ | $a_1$ | $a_2$ | $b_1$ | $b_2$ |
| *HighLow* | Coef. | -0.317 | 0.480 | -0.036 | 0.549 | -0.459 | 0.457 | -0.124 | 0.479 |
| | p-value | (<.001) | (<.001) | (0.519) | (<.001) | (<.001) | (<.001) | (0.046) | (<.001) |
| *Spread* | Coef. | -0.471 | 0.619 | -0.026 | 0.551 | -0.519 | 0.554 | -0.112 | 0.475 |
| | p-value | (<.001) | (<.001) | (0.647) | (<.001) | (<.001) | (<.001) | (0.116) | (<.001) |
| *EffSprd* | Coef. | -0.181 | 0.401 | -0.025 | 0.554 | -0.112 | 0.240 | -0.150 | 0.502 |
| | p-value | (0.003) | (<.001) | (0.808) | (<.001) | (0.016) | (<.001) | (<.001) | (<.001) |
| *NearDepth* | Coef. | 0.443 | -0.215 | 0.081 | 0.543 | 0.652 | -0.142 | 0.350 | 0.376 |
| | p-value | (<.001) | (<.001) | (0.407) | (<.001) | (<.001) | (<.001) | (0.014) | (<.001) |

Panel B: Cross-Sectional Median Coefficient Estimate when Instruments are *SpreadNotNAS* and *RunsNotI*

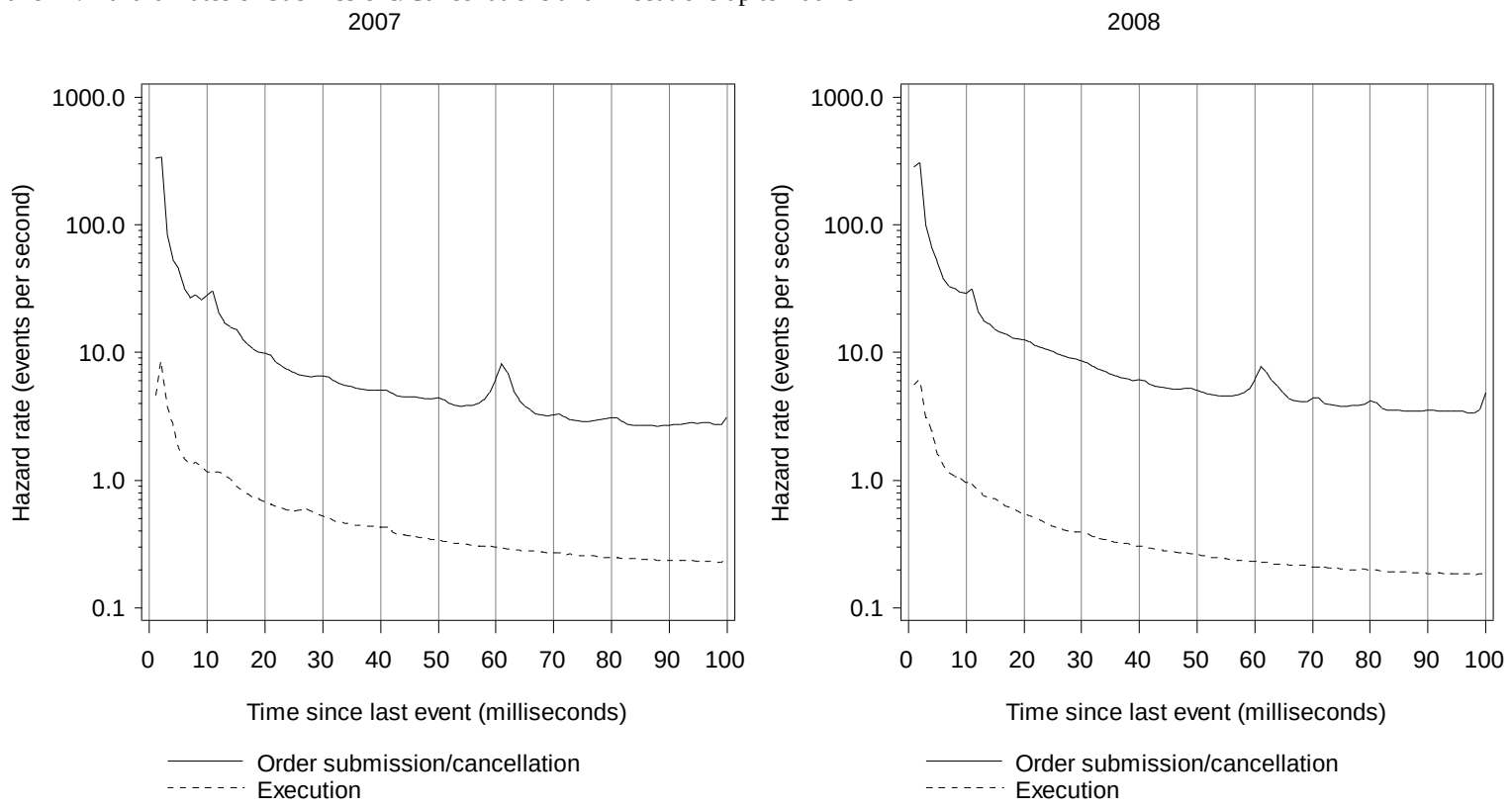
| | | 2007 | | | | 2008 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $a_1$ | $a_2$ | $b_1$ | $b_2$ | $a_1$ | $a_2$ | $b_1$ | $b_2$ |
| *HighLow* | Coef. | -0.331 | 0.390 | -0.119 | 0.511 | -0.398 | 0.414 | -0.157 | 0.461 |
| | p-value | (<.001) | (<.001) | (0.112) | (<.001) | (<.001) | (<.001) | (0.030) | (<.001) |
| *Spread* | Coef. | -0.214 | 0.790 | -0.065 | 0.534 | -0.132 | 0.842 | -0.077 | 0.485 |
| | p-value | (<.001) | (<.001) | (0.114) | (<.001) | (<.001) | (<.001) | (0.091) | (<.001) |
| *EffSprd* | Coef. | -0.149 | 0.341 | -0.129 | 0.524 | -0.064 | 0.227 | -0.131 | 0.516 |
| | p-value | (0.087) | (<.001) | (0.172) | (<.001) | (0.625) | (<.001) | (0.085) | (<.001) |
| *NearDepth* | Coef. | 0.325 | -0.299 | 0.175 | 0.503 | 0.568 | -0.204 | 0.309 | 0.392 |
| | p-value | (<.001) | (<.001) | (0.532) | (<.001) | (<.001) | (<.001) | (<.001) | (0.016) |

# Figure 1

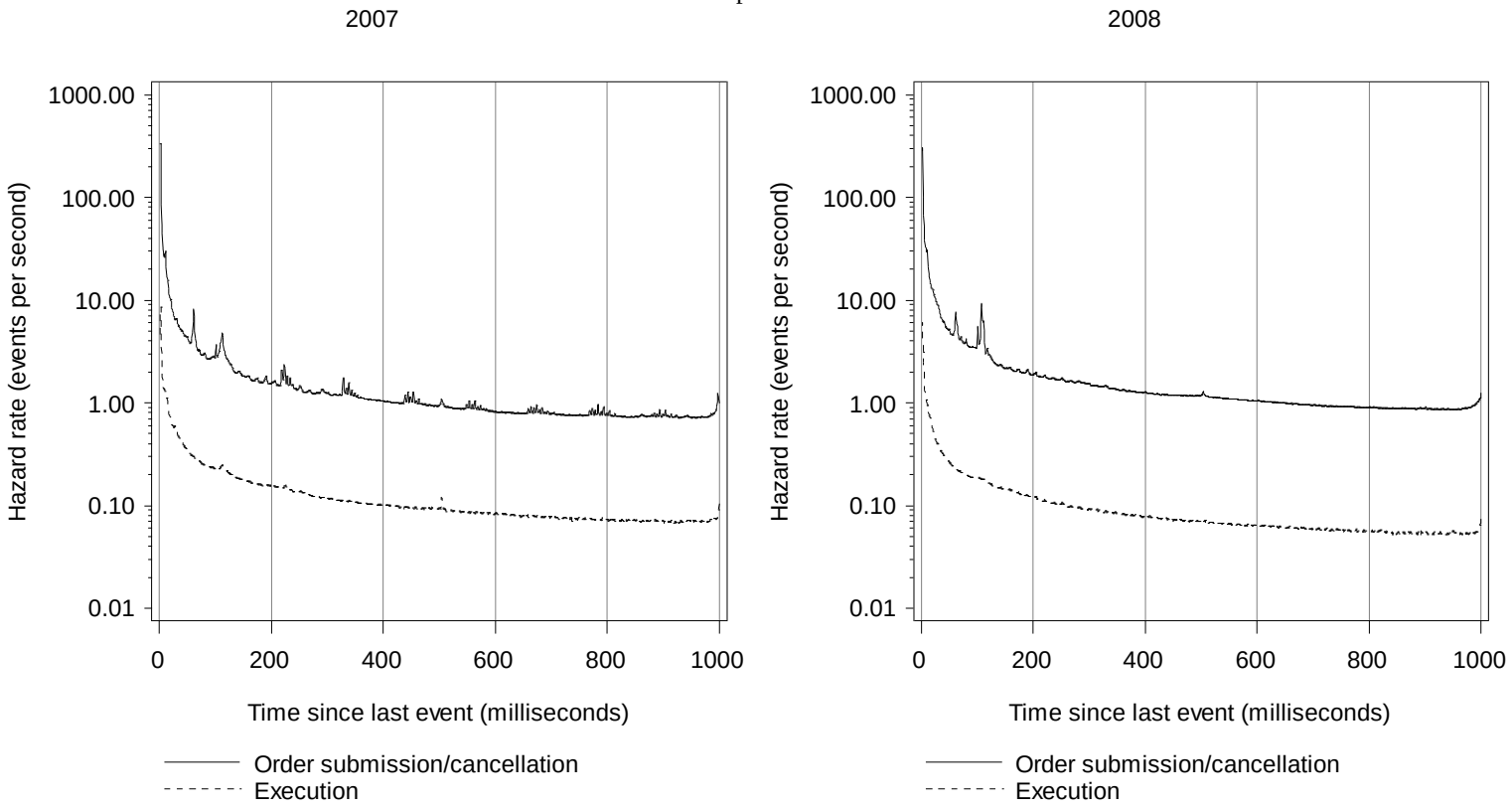
## Hazard Rates of Orders and Trades

This figure presents estimated hazard rates for (i) order submissions and cancellations (i.e., all messages that do not involve trade execution), and (ii) trade executions. In the estimation of the submission/cancellation hazard rate, execution is assumed to be an exogenous censoring process, while in the estimation of the execution hazard rate, submissions and cancellations are assumed to be the exogenous censoring process. The estimated hazard rate plotted at time *t* is the estimated average over the interval [*t*–1 ms, *t*). The hazard rate for submissions/cancellations can be interpreted as the intensity of submissions and cancellations of limit orders conditional on the elapsed time since any market event (which can be a submission, a cancellation, or an execution). Similarly, the hazard rate for execution of trades can be interpreted as the intensity of executions conditional on the elapsed time subsequent to any market event. The hazard rates are estimated using the life-table method. In Panel A, we plot the hazard rates up to 100 milliseconds side-by-side for the 2007 and 2008 sample periods. This plot enables us to observe in greater detail very short-term patterns. In Panel B we plot the hazard rates up to one second.

Panel A: Hazard Rates of Submissions/Cancellations and Executions up to 100ms



Panel B: Hazard Rates of Submissions/Cancellations and Executions up to 1000ms

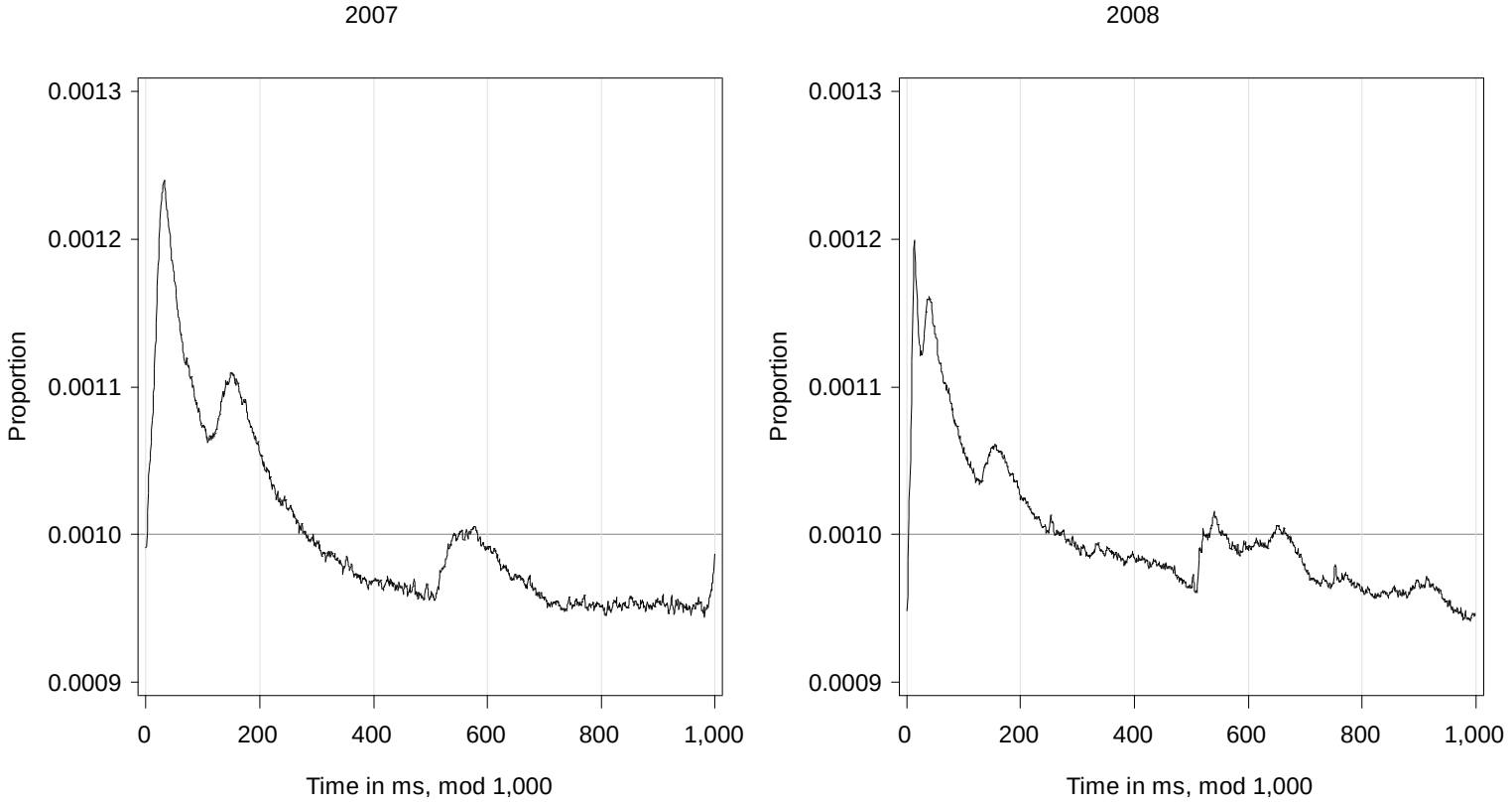


# Figure 2

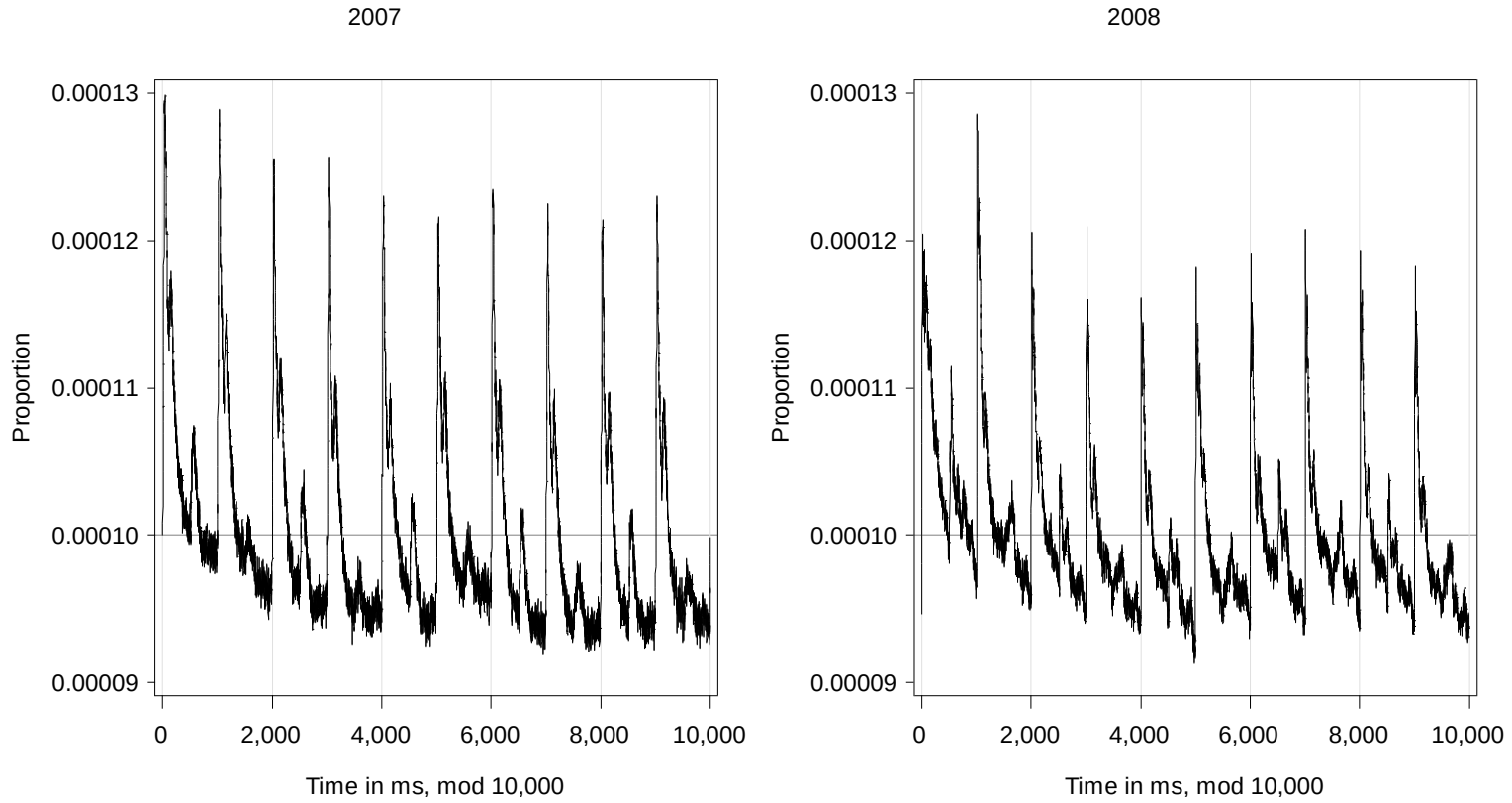
## Clock-time Periodicities of Market Activity

This figure presents clock-time periodicities in message arrival to the market. The original time stamps are milliseconds past midnight. The one-second remainder is the time stamp mod 1,000, i.e., the number of milliseconds past the one-second mark. In Panel A, we plot the sample distribution of one-second remainders side-by-side for the 2007 and 2008 sample periods. The ten-second remainder is the time stamp mod 10,000, the number of milliseconds past the ten-second mark. Panel B plots the sample distribution of ten-second remainders. The horizontal lines in the graphs indicate the position of the uniform distribution (the null hypothesis).

Panel A: Sample Distributions of One-Second Millisecond Remainders



Panel B: Sample Distributions of Ten-Second Millisecond Remainders

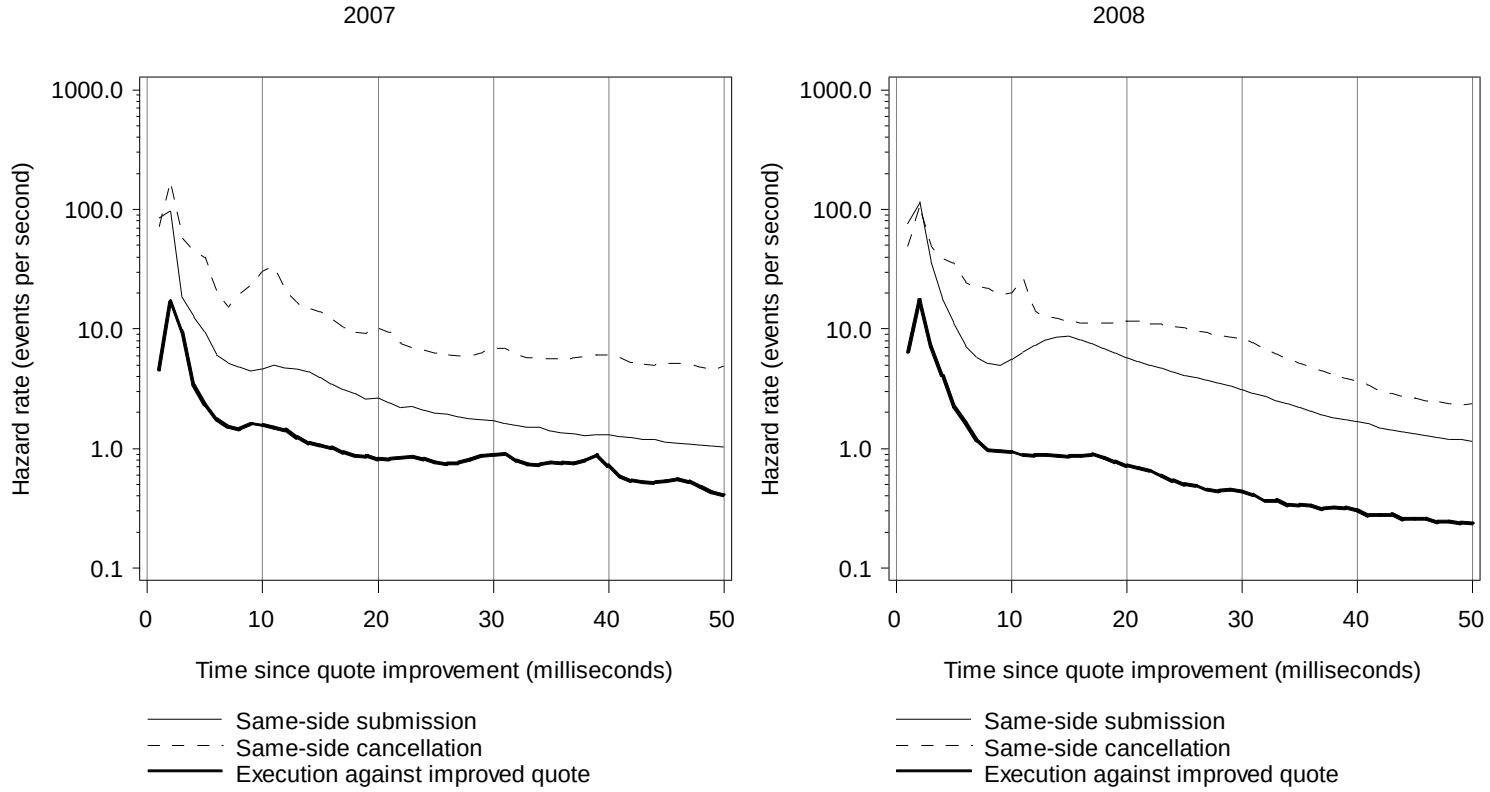


# Figure 3

## Speed of Response to Market Events

This figure looks at the speed of responses to certain market events that have well-defined economic meaning. In Panel A, the market event is an improved quote via the submission of a new limit order—either an increase in the best bid price or a decrease in the best ask price. Subsequent to this market event, we estimate (separately) the hazard rates for three types of responses: (i) a limit order submission on the same side as the improvement (e.g., buy order submitted following an improvement in the bid price), (ii) a cancellation of a standing limit order on the same side, and (iii) an execution against the improved quote (e.g., the best bid price is executed by an incoming sell order). In Panel B, the market event is deterioration in the quote as a result of a cancellation of a standing limit order (e.g., a limit buy order alone at the best bid price is cancelled and the best bid price therefore decreases). Subsequent to this market event, we estimate (separately) the hazard rates for three types of responses: (i) a limit order submission on the same side as the quote deterioration, (ii) a cancellation of a standing limit order on the same side, and (iii) an execution against the worsened quote. In all estimations, any event other than the one whose hazard rate is being estimated is taken as an exogenous censoring event. The estimated hazard rate plotted at time *t* is the estimated average over the interval [*t*–1 ms, *t*). The hazard rate for a response can be interpreted as the intensity of the response conditional on the elapsed time since the conditioning market event (e.g., the improved quote in Panel A).

Panel A: Responses to Quote Improvement



Panel B: Responses to Quote Deterioration Due to a Limit Order Cancellation