Available online at www.sciencedirect.com

ScienceDirect

Journal of Financial Markets 16 (2013) 646–679

Journal of FINANCIAL MARKETS

www.elsevier.com/locate/finmar

Low-latency trading ☆

Joel Hasbrouck[a,*], Gideon Saar[b,1]

[a]Stern School of Business, 44 West 4th Street, New York, NY 10012, USA
[b]Johnson Graduate School of Management, Cornell University, 455 Sage Hall, Ithaca, NY 14853, USA

Received 16 January 2013; received in revised form 13 May 2013; accepted 13 May 2013
Available online 22 May 2013

Abstract

We define low-latency activity as strategies that respond to market events in the millisecond environment, the hallmark of proprietary trading by high-frequency traders though it could include other algorithmic activity as well. We propose a new measure of low-latency activity to investigate the impact of high-frequency trading on the market environment. Our measure is highly correlated with NASDAQ-constructed estimates of high-frequency trading, but it can be computed from widely-available message data. We use this measure to study how low-latency activity affects market quality both during normal market conditions and during a period of declining prices and heightened economic uncertainty. Our analysis suggests that increased low-latency activity improves traditional market quality measures—decreasing spreads,

☆We are grateful for comments from Tarun Chordia (the editor), Charles Jones, Andrew Karolyi, Albert Menkveld, Ciamac Moallemi, Maureen O'Hara, Harvey Westbrook, an anonymous referee, and seminar (or conference) participants at the Chicago Quantitative Alliance/ Society of Quantitative Analysts, the Conference on Current Issues in Financial Regulation (University of Notre Dame), Cornell's Johnson School, Cornell Financial Engineering Manhattan, the CREATES Market Microstructure Symposium (Aarhus), Erasmus University, ESSEC Business School, Humbolt University, the Investment Industry Regulatory Organization of Canada/ DeGroote School, the National Bureau of Economic Research Market Microstructure Group meeting, New York University, Rutgers Business School, SAC Capital, University of Toronto, the Western Finance Association meetings, and the World Federation of Exchanges Statistics Advisory Group. DISCLAIMER: This research was not specifically supported or funded by any organization. During the period over which this research was developed, Joel Hasbrouck taught (for compensation) in the training program of a firm that engages in high-frequency trading, and served as a member (uncompensated) of a CFTC advisory committee on high-frequency trading. Gideon Saar serves as a member (uncompensated) of FINRA's Economic Advisory Committee. This paper uses NASDAQ OMX ITCH data that are generally available by paid subscription to practitioners but are generally made available at no cost to academics. One section of this paper also uses data on high-frequency trading generally provided to academics at no cost by NASDAQ OMX.

*Corresponding author. Tel.: +1 212 998 0310; fax: +1 212 995 4233.
E-mail addresses: jhasbrou@stern.nyu.edu (J. Hasbrouck), gs25@cornell.edu (G. Saar).
[1]Tel.: +1 607 255 7484.

1386-4181/$ - see front matter
http://dx.doi.org/10.1016/j.finmar.2013.05.003

increasing displayed depth in the limit order book, and lowering short-term volatility. Our findings suggest that given the current market structure for U.S. equities, increased low-latency activity need not work to the detriment of long-term investors.

JEL classification: G10; G20; G23; G28

Keywords: High-frequency trading; Limit order markets; NASDAQ; Order placement strategies; Liquidity; Market quality

1. Introduction

Our financial environment is characterized by an ever increasing pace of both information gathering and the actions prompted by this information. Speed in absolute terms is important to traders due to the inherent fundamental volatility of financial securities. Relative speed, in the sense of being faster than other traders, is also very important because it can create profit opportunities by enabling a prompt response to news or market activity. 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. 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 in close proximity 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 couple of milliseconds.

The beneficiaries from this massive investment in technology appear to be a new breed of high-frequency traders 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 serving as the intermediaries, e.g., Menkveld (in this issue)], it is unclear whether intense low-latency activity harms or helps the market.

Our goal in this paper is to examine the influence of these low-latency traders on certain dimensions of market quality. More specifically, we would like to know how their combined activity affects attributes such as bid-ask spreads, the total price impact of trades, depth in the limit order book, and the short-term volatility of stocks.[3] To investigate these questions, we utilize publicly-available NASDAQ order-level data that are identical to those supplied to subscribers and provide real-time information about orders and executions on NASDAQ. Each

[2]See, for example, the discussion of high-frequency traders in the SEC's Concept Release on Equity Market Structure (2010).

[3]Another dimension of market quality, the informational efficiency of prices (or price discovery), and its relationship to high-frequency trading is investigated in Brogaard, Hendershott, and Riordan (2012), and Carrion (in this issue).

entry (submission, cancellation, or execution) is time-stamped to the millisecond, and hence these data provide a very detailed view of NASDAQ activity.

We begin by providing a discussion of the players in this new millisecond environment: proprietary and agency algorithms. We document periodicities in the time-series of market activity, which we attribute to agency algorithms. We also look at the speed at which some traders respond to market events—the hallmark of proprietary trading by high-frequency trading firms—and find that the fastest traders have an effective latency of 2–3 ms during our sample period.

We propose a new measure of low-latency activity based on "strategic runs" of linked messages that describe dynamic order placement strategies. While our measure might reflect some activity originating from agency algorithms, our restriction to long strategic runs makes it more likely that the measure predominately captures the activity of high-frequency traders, and we believe that it is highly correlated with their presence in the market. As such, we view this measure as a proxy for the activity of high-frequency traders. An advantage of our measure is that it can be constructed from publicly-available data, and therefore does not rely on specialty datasets that may be limited in scale and scope. We show that our measure is highly correlated with aggregate trading by high-frequency trading firms in the 120-stock NASDAQ HFT dataset studied in Brogaard (2012), Brogaard, Hendershott, and Riordan (2012), and Carrion (in this issue). To assess robustness, we attempt to exclude agency algorithms from our measure, and find that our conclusions are unchanged. However, due to the manner in which the measure is constructed, there is no certainty that it only captures high-frequency trading.

We use our measure to examine how the intensity of low-latency activity affects several market quality measures. We find that an increase in low-latency activity reduces quoted spreads and the total price impact of trades, increases depth in the limit order book, and lowers short-term volatility. Our results suggest that the increased activity of low-latency traders is beneficial to traditional benchmarks of market quality in the current U.S. equity market structure, one that is characterized by both high fragmentation and wide usage of agency and proprietary algorithms. We use a variety of econometric specifications to examine the robustness of our conclusions.

Furthermore, we employ two distinct sample periods to investigate whether the impact of low-latency trading on market quality differs between normal periods and those associated with declining prices and heightened uncertainty. Over October 2007, our first sample period, stock prices were relatively flat or slightly increasing. Over our second sample period, June 2008, stock prices declined (the NASDAQ index was down 8% in that month) and uncertainty was high following the fire sale of Bear Stearns. We find that higher low-latency activity enhances market quality in both periods.[4]

Our paper relates to small but growing strands in the empirical literature on speed in financial markets and high-frequency trading (which is a subset of algorithmic trading comprised of proprietary algorithms that require low latency). With regard to speed, Hendershott and Moulton (2011) and Riordan and Storkenmaier (2012) examine market-wide changes in technology that reduce the latency of 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 (Prix, Loistl, and Huetl, 2007; Groth, 2009; Gsell, 2009; Gsell and Gomber, 2009; Hendershott and Riordan,

[4]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.

forthcoming), the interdealer foreign exchange market (Chaboud, Hjalmarsson, Vega and Chiquoine, 2013), and the U.S. equity market (Hendershott, Jones, and Menkveld, 2011).

A smaller set of papers focuses on high-frequency trading. Kirilenko, Kyle, Samadi, and Tuzun (2011) look at high-frequency traders in the futures market during the flash crash episode. Brogaard (2012) seeks to characterize high-frequency trading on NASDAQ and BATS, while Brogaard, Hendershott, and Riordan (2012) study the impact of high-frequency trading on price discovery in U.S. equities. Three other papers also appear in this special issue on high-frequency trading. Menkveld (in this issue) is a case study of a particular high-frequency trader who acts as a market maker on Chi-X and Euronext. Carrion (in this issue) uses the NASDAQ HFT dataset to examine the sources of profitability of high-frequency trading firms, how they carry out their strategies, and their impact on market efficiency. Hagströmer and Norden (in this issue) use special data from NASDAQ OMX Stockholm to separately characterize the strategies of "market making" and "opportunistic" high-frequency trading firms.

The rest of this paper proceeds as follows. The next section describes our sample and data. Section 3 provides an introductory discussion of the millisecond environment with some evidence on the activity of proprietary and agency algorithms. Section 4 describes our measure of low-latency activity. In Section 5 we estimate the impact of our measure on diverse measures of market quality. In Section 6 we discuss related papers and place our findings within the context of the literature, and Section 7 concludes.

2. Data and sample

2.1. NASDAQ order-level data

The NASDAQ Stock Market 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 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 publicly-available 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 observe four different types of messages: (1) the addition of a displayed order to the book, (2) the cancellation (or partial cancellation) of a displayed order, (3) the execution (or partial execution) of a displayed order, and (4) the execution (or partial execution) of a non-displayed order. In other words, we observe every displayed order that arrives to the NASDAQ market, including the NASDAQ portion of Reg NMS Intermarket Sweep Orders and odd-lot orders. We do not observe submission and cancellation of non-displayed non-marketable limit orders, which are unobservable to market participants in real-time and hence are not part of the TotalView-ITCH data feed. Since we

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

observe all trades (including odd-lots), however, we know when a non-displayed limit order is executed.[6]

2.2. 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.[7] We then take the top 500 stocks, ranked by market capitalization as of September 30, 2007. Our first sample period is October 2007 (23 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 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 period, 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 10% of the 10-minute intervals have fewer than 250 messages. Net of these exclusions, the 2007 sample contains 351 stocks, and the 2008 sample contains 399 stocks. Our results concerning the impact of low-latency activity on market quality are robust to imposing a less stringent screen that leaves more than 90% of the stocks in the sample.[8]

Table 1 provides summary statistics for the stocks in both sample periods using information from CRSP and the NASDAQ data. 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 $635 with a median of $29. Panel B summarizes data collected from NASDAQ. In 2007, the median firm had 26,862 limit order submissions (daily average), 24,015 limit order cancellations, and 2,482 marketable order executions.[9]

[6]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.

[7]NASDASQ introduced the three-tier initiative for listed stocks in July 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).

[8]Specifically, the less stringent screen only excludes stocks if more than 10% of the 10-minute intervals have fewer than 100 messages. This screen significantly increases the number of stocks in both sample periods (471 in 2007 and 456 in 2008), but the results are very similar to those discussed in Section 5 and presented in Tables 5–7.

[9]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 standing limit orders is 45,508–40,943=4,565. This is above the reported mean number of marketable orders executed (3,791) because a single marketable order may involve multiple standing limit orders. As we describe in footnote 6, we group executions of standing limit orders that were triggered by a single marketable order into one event.

Table 1
Summary statistics.

The sample consists of the 500 largest firms (as ranked by market capitalization as of September 28, 2007) over two periods: October 2007 (23 trading days), and June 2008 (21 trading days). A firm is dropped from the sample if the proportion of 10-minute intervals with fewer than 250 messages is above 10%. After applying the screen, our sample consists of 351 stocks in the October 2007 sample period and 399 stocks in the June 2008 sample period. Panel A reports statistics derived from the CRSP database. Equity market capitalization is as of the last trading day prior to the start of the sample period. Panel B presents statistics derived from the NASDAQ TotalView-ITCH database. Depth on the book, near depth (within 10 cents of the best bid or ask) and quoted spread are time-weighted averages for each firm. The effective spread is defined as the twice the transaction price less the quote midpoint for a marketable buy order (or twice the midpoint less the transaction price for a sell order), and the average is share-weighted.

Panel A. CRSP summary statistics

	2007				2008			
	Market capitalization ($Million)	Avg. closing Price ($)	Avg. daily volume (1,000s)	Avg. daily return (%)	Market capitalization ($Million)	Avg. closing price ($)	Avg. daily volume (1,000s)	Avg. daily return (%)
Mean	6,609	37.09	3,172	0.109	5,622	31.88	2,931	−0.565
Median	2,054	29.08	1,074	0.130	1,641	24.96	1,111	−0.516
Std	20,609	41.54	8,083	0.570	19,348	38.93	6,410	0.615
Min	789	2.22	202	−2.675	286	2.32	112	−3.449
Max	275,598	635.39	77,151	1.933	263,752	556.32	74,514	0.817

Panel B. NASDAQ (TotalView-ITCH) summary statistics

		Avg. daily limit order submissions	Avg. daily limit order cancellations	Avg. daily marketable order executions	Avg. daily shares executed (1,000s)	Average depth (1,000s)	Average near depth (1,000s)	Average quoted spread ($)	Average effective spread ($)
2007	Mean	45,508	40,943	3,791	1,400	486	57	0.034	0.025
	Median	26,862	24,015	2,482	548	147	11	0.025	0.019
	Std	73,705	68,204	4,630	3,231	1,616	257	0.032	0.021
	Min	9,658	8,013	695	130	26	1	0.010	0.009
	Max	985,779	905,629	62,216	32,305	15,958	3,110	0.313	0.214
2008	Mean	54,287	50,040	3,694	1,203	511	43	0.035	0.023
	Median	34,658	31,426	2,325	483	154	10	0.023	0.016
	Std	61,810	56,728	4,676	2,618	1,767	152	0.041	0.024
	Min	8,889	7,983	291	42	20	1	0.010	0.008
	Max	593,143	525,346	61,013	32,406	25,004	2,482	0.462	0.257

3. The millisecond environment

Much trading and message activity in U.S. equity markets is commonly attributed to trading algorithms.[10] 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, we can categorize algorithmic activity as proprietary or agency. We consider high-frequency trading a subcategory of proprietary algorithms for which low latency is essential. Our paper mostly focuses on this low-latency activity and its impact on the market, but to establish the context we discuss both agency and proprietary algorithms in this section.

Agency algorithms are used by buy-side institutions (and the brokers who serve them) to minimize the cost of executing trades in the process of implementing changes in their investment portfolios. They have been in existence for about two decades, but the last 10 years have witnessed a dramatic increase in their appeal due to decimalization (in 2001) and increased fragmentation in the 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. 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 their ultimate goal is to execute a desired position change. Hence they essentially demand liquidity, even though their strategies might utilize nonmarketable limit orders.

In terms of technological requirements, agency algorithms are mostly based on historical estimates of price impact and execution probabilities across multiple trading venues and over time, and often do not require much real-time input except for tracking the pieces of the orders they execute. For example, volume-weighted average price 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 agency algorithms drive one of the most curious patterns we observe in the millisecond environment: clock-time periodicity. For a given timestamp t, the quantity mod $(t,1000)$ is the millisecond remainder, i.e., a millisecond time stamp 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.

Fig. 1 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: large peaks occurring shortly after the one-second boundary at roughly 10–30 ms and around 150 ms, as well as broad elevations around 600 ms. We believe that these peaks are indicative of agency algorithms that simply check market conditions and

[10]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.

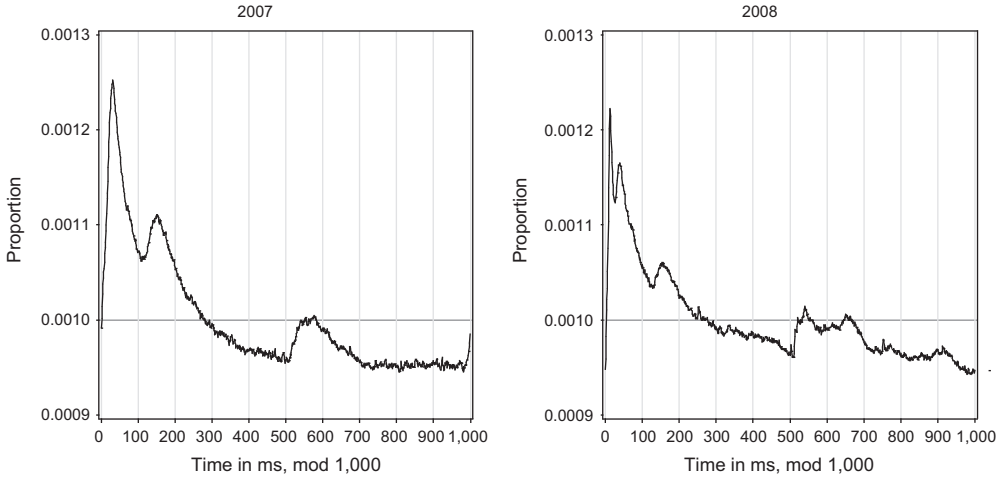

Fig. 1. Clock-time periodicities of market activity. This figure presents clock-time periodicities in message arrival to the market. The data contain millisecond time stamps. The one-second remainder is the time stamp mod 1,000, i.e., the number of ms past the one-second mark. We plot the sample distribution of one-second remainders side-by-side for the 2007 and 2008 sample periods. The horizontal lines in the graphs indicate the position of the uniform distribution (the null hypothesis).

execution status every second (or minute), near the second (or the half-second) boundary, and respond to the changes they encounter. These periodic checks are subject to latency delays (i.e., 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 the algorithm: 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) or technology.[11]

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. However, as long as someone is sending messages in a periodic manner, strategic responses by others who monitor the market continuously could serve to amplify rather than eliminate the periodicity. The clustering of agency algorithms means that the provision of liquidity by proprietary algorithms or by one investor to another is higher at these times, and hence conceivably helps agency algorithms execute their orders by increasing available liquidity. As such, agency algorithms would have little incentive to change, making these patterns we identify in the data persist over time.[12] It is also possible, however, that the major players in the industry that designs and implements agency algorithms were unaware of the periodicity prior to our research. If this is indeed the case, and the predictability of buy-side order flow is considered

[11]We checked with NASDAQ whether their systems that provide traders with more complex order types (e.g., RASH) could be the source of these clock-time periodicities. NASDAQ officials contend that their systems do not create such periodicities.

[12]This intuition is similar in spirit to Admati and Pfleiderer (1988), where uninformed traders choose to concentrate their trading at certain times in order to gain from increased liquidity even in the presence of informed traders.

undesirable for various reasons, our findings could lead to changes in the design of agency algorithms that would eliminate such periodicities in the future.

Relative to agency algorithms, proprietary algorithms are more diverse and more difficult to concisely characterize. Nonetheless, our primary focus is a new breed of proprietary algorithms that utilizes extremely rapid response to the market environment. Such algorithms, which are meant to profit from the trading environment itself (as opposed to investing in stocks), are employed by hedge funds, proprietary trading desks of large financial firms, and independent specialty firms. These algorithms can be used, for example, to provide liquidity or to identify a trading interest in the market and use that knowledge to generate profit. Brogaard (2012) and Brogaard, Hendershott, and Riordan (2012) study a 120-stock dataset in which NASDAQ identified the trading by 26 high-frequency firms in 2008 and 2009. They report that these firms are involved in 68.5% of NASDAQ dollar volume traded over that time period.[13]

The hallmark of high-frequency proprietary algorithms is speed: low-latency capabilities. These traders invest in co-location and advanced computing technology to create an edge in strategic interactions. Their need to respond to market events distinguishes them from the majority of agency algorithms. We define low-latency trading as "strategies that respond to market events in the millisecond environment." This definition is meant to capture all proprietary algorithms that require low latency (i.e., high-frequency traders) but could potentially include some agency algorithms that utilize low-latency capabilities. How fast are the low-latency traders? The definition above, which is formulated in terms of speed of response to market events, suggests that an answer to this question could be found by focusing on market 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 new 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. Events on the sell side of the book, subsequent to a decrease in the ask price, can be defined in a similar fashion.

We therefore estimate the hazard rates (i.e., the message arrival intensities) of the above specific responses subsequent to order submissions that improve the quote. In Fig. 2 we plot separately the conditional hazard rates for same-side submissions, same-side cancellations, and executions against the improved quotes (pooled over bid increases and ask decreases). We observe pronounced peaks at approximately 2–3 ms, particularly for executions. This suggests that the fastest responders—the low-latency traders—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 ms (Kosinski, 2012). The figure suggests that the time it takes for some low-latency traders to observe a market event, process the information, and act on it is indeed very short.

Since humans cannot follow such low-latency activity on their trading screens, one might wonder what it actually looks like. 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 34 submissions (and 33 cancellations) of orders to buy 300 shares. 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. We highlight in gray some of the

[13]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 strategies may be somewhat larger.

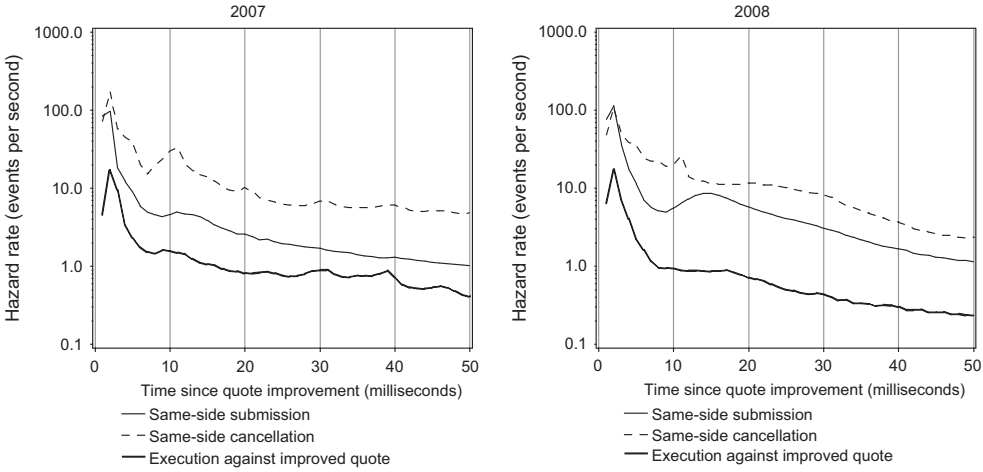

Fig. 2. Speed of response to market events. This figure depicts response speeds subsequent to a specific market event. The market event is an improved quote via the submission of a new limit order—either an increase in the best bid price or a decrease in the best ask price. Subsequent to this market event, we estimate (separately) the hazard rates for three types of responses: a limit order submission on the same side as the improvement (e.g., buy order submitted following an improvement in the bid price); a cancellation of a standing limit order on the same side; and, an execution against the improved quote (e.g., the best bid price is executed by an incoming sell order). 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 event.

orders and cancellations in the table to make it easier to see what appear to be two algorithms that are engaged in strategic behavior attempting to position themselves at the top of the book: undercutting each other, canceling and resubmitting when the other algorithm cancels, and so on. The pricing of the orders causes the bid quote to rapidly oscillate between $20.04 and $20.05. 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 of Table 2 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 could be 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 create their own space in the sense that some of what they do seems to be intended to trigger a response from (or respond to) other algorithms. Activity in the limit order book is dominated by the interaction among automated algorithms, in contrast to a decade ago when human traders still ruled.

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

Table 2
Examples of strategic runs for ticker symbol ADCT on October 2, 2007.

A strategic run is a series of submissions, cancellations, and executions that are linked by direction, size, and timing, and which are likely to arise from a single algorithm. The examples are taken from activity in one stock (ADC Telecommunications, ticker symbol ADCT) on October 2, 2007. In the two cases presented here, the activity constitutes all messages in this stock. Panel A reports order activity starting around 9:51:57 am. Shading identifies messages corresponding to 100- and 300-share runs. Panel B reports activity starting around 9:57:18 am. This run is active for 1 minute and 18 seconds, and comprises 142 messages.

Panel A: ADCT order activity starting at 9:51:57.849

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 at 9:57:18.839

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.						

(e.g., diversification), the social benefits of high-frequency proprietary trading are more elusive. If high-frequency proprietary algorithms engage in electronic liquidity provision, then they provide a similar service to that of traditional market makers, bridging the intertemporal disaggregation of order flow in continuous markets. However, the social benefits of other types of low-latency trading are more difficult to ascertain. One argument sometimes made in the context of proprietary statistical arbitrage algorithms is that they aid price discovery by eliminating transient price disturbances, but such an argument in a millisecond environment is tenuous: at such speeds and in such short intervals it is difficult to determine the price component that constitutes a real innovation to the true value of a security as opposed to a transitory influence. The social utility in algorithms that identify buy-side interest and trade ahead of it is even harder to defend. It therefore becomes an empirical question to determine whether these high-frequency trading algorithms in the aggregate harm or improve the market quality perceived by long-term investors. Our paper seeks to answer this question.

4. Low-latency trading and market quality: the measures

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.[14] 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 could be measured using time intervals that are easily observed by these investors. Therefore, in this section we develop measures that would allow us to characterize the influence of low-latency trading on liquidity and short-term volatility observed over 10-minute intervals throughout the day.

To construct a measure of low-latency activity, we begin by identifying "strategic runs," which are linked submissions, cancellations, and executions that are likely to be parts of a dynamic algorithmic strategy. Our goal is to isolate instances of market activity that look like the interactions presented in Table 2. 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 100 ms 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% of the cancellations in the 2007 sample, and 54% in the 2008 sample. Although we allow up to 100 ms to elapse from cancellation to resubmission, 49% of the imputed durations are one or zero ms, and less than 10% exceed 40 ms. The length of a run

[14]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).

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. We identify about 46.0 million runs in the 2007 sample period and 67.1 million runs in the 2008 sample period.

Table 3 presents summary statistics for the runs. We observe that around 75% of the runs have 3 to 9 messages, but longer runs (10 or more messages) constitute over 60% of the messages that are associated with strategic runs. The proportion of runs that are (at least partially) executed is 38.1% in 2007 and 30.5% in 2008. About 8.1% (7.1%) 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.

To construct a measure of low-latency trading that is more robust to measurement error, we transform the raw strategic runs in two ways. The first transformation is to use only longer runs —runs of 10 or more messages—to construct the measure. While our methodology to impute links between cancellations and resubmissions of orders can result in misclassifications, for a run with many resubmissions to arise solely as an artifact of such errors 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. While the 10-message cutoff is somewhat arbitrary, these runs represent more than 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. Our conclusions on the impact of low-latency activity on market quality are unchanged when we include all runs.[15]

The second transformation we use to reduce measurement error is to utilize time-weighting of the number of runs rather than simply aggregating the runs or the messages in runs. We define our measure of low-latency activity, *RunsInProcess*, as the time-weighted average of the number of strategic runs of 10 messages or more the stock experiences in an interval.[16] Time-weighting helps us combat potential errors because it ensures that roughly equivalent patterns of activity contribute equally to our measure, which can be demonstrated using the strategic run shown in Panel B of Table 2. This run, which lasts 78.5 seconds, contributes 0.129 (78.5/600) to *RunsInProcess* of stock ADCT in the interval 9:50–10:00 am on October 2, 2007. What if we were wrong and the inferred resubmission at time 9:57:20.761 actually came from a different algorithm, so that the activity described in Panel B of Table 2 was generated by one 48-message algorithm and another 94-message algorithm rather than a single 142-message algorithm? This should not alter our inference about the activity of low-latency traders from an economic standpoint, because the two shorter algorithms together constitute almost the same amount of low-latency activity as the single longer algorithm. The time-weighting of *RunsInProcess*

[15]To ensure that omitting shorter runs does not materially affect our conclusions, we used all strategic runs to construct an alternative measure of low-latency activity: *AllRunsInProcess*, and carried out exactly the same analysis. The results were similar to those discussed in Section 5 and presented in Tables 5–7.

[16]The time-weighting of this measure works as follows. Suppose we construct this variable for the interval 9:50:00 am–10:00:00 am. If a strategic run started at 9:45:00 am and ended at 10:01:00 am, it was active for the entire interval and hence it adds 1 to the *RunsInProcess* measure. A run that started at 9:45:00 am and ended at 9:51:00 am 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.

Table 3
Strategic runs.

A strategic run is a series of submissions, cancellations, and executions that are linked by direction, size, and timing, and which are likely to arise from a single algorithm. We sort runs into categories by length (i.e., the number of linked messages). The table reports statistics on number of runs, messages, and executions (separately active and passive) within each category. An active execution occurs when a run is terminated by canceling a resting limit order and submitting a marketable limit order. A passive execution occurs when a resting limit order that is part of a run is (at least in part) executed by an incoming marketable order.

	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	20,294,968	44.11%	79,695,563	15.67%	1,954,468	9.63%	4,981,521	24.55%	6,922,605	34.11%
	5–9	13,540,437	29.43%	89,204,570	17.54%	1,012,573	7.48%	4,715,922	34.83%	5,706,905	42.15%
	10–14	5,650,415	12.28%	65,294,103	12.84%	267,517	4.73%	1,808,138	32.00%	2,069,393	36.62%
	15–19	1,854,002	4.03%	31,229,102	6.14%	153,839	8.30%	654,241	35.29%	805,414	43.44%
	20–99	4,337,029	9.43%	153,384,374	30.16%	301,266	6.95%	1,575,876	36.34%	1,871,244	43.15%
	100+	333,308	0.72%	89,735,209	17.65%	26,039	7.81%	116,465	34.94%	141,962	42.59%
	All	46,010,159	100.00%	508,542,921	100.00%	3,715,702	8.08%	13,852,163	30.11%	17,517,523	38.07%
2008	3–4	31,012,203	46.24%	122,325,313	19.53%	2,427,326	7.83%	5,552,338	17.90%	7,970,158	25.70%
	5–9	19,758,076	29.46%	130,370,772	20.82%	1,287,276	6.52%	5,436,189	27.51%	6,705,727	33.94%
	10–14	7,941,089	11.84%	91,486,978	14.61%	385,902	4.86%	2,186,628	27.54%	2,566,974	32.33%
	15–19	2,533,217	3.78%	42,663,802	6.81%	219,403	8.66%	795,483	31.40%	1,012,340	39.96%
	20–99	5,583,768	8.33%	191,395,420	30.56%	398,771	7.14%	1,712,015	30.66%	2,105,346	37.70%
	100+	239,751	0.36%	48,084,901	7.68%	15,541	6.48%	62,838	26.21%	78,171	32.61%
	All	67,068,104	100.00%	626,327,186	100.00%	4,734,219	7.06%	15,745,491	23.48%	20,438,716	30.47%

ensures that the measure computed from the two algorithms is almost identical to the one originally computed from the single algorithm (the two will differ only by 0.005/600 = 0.000008 due to the 5 ms gap between the end of the first algorithm and the beginning of the second algorithm), and hence this type of error would not affect our empirical analysis.

It is important to recognize that our measure of low-latency activity does not have an inherently positive relationship with market quality. In fact, if liquidity is provided by patient limit order traders (which is the case most often described in theoretical models), depth in the book is maximized when the cancellation rate is zero. In other words, liquidity is highest when limit orders stay in the book until they are executed, in which case our measure *RunsInProcess* is equal to zero. As traders begin cancelling orders, liquidity in the book worsens and our measure increases. This suggest that holding everything else equal, *RunsInProcess* should be negatively related to liquidity, though liquidity may decline only modestly if traders cancel but replace limit orders with other limit orders rather than switch to marketable orders. However, the relationship between *RunsInProcess* and liquidity is more complex because low-latency traders may be willing to submit more limit orders and provide more depth if they have the technology to cancel limit orders quickly enough to lower the pick-off risk of their orders. Hence, we do not know a priori whether the relationship between our measure of low-latency activity and market quality is positive or negative in equilibrium, and this is what we test in Section 5.

Our measure captures low-latency activity.[17] One definition of "high-frequency trading" is proprietary trading that utilizes low latency. Trading firms that use low-latency technology include all high-frequency traders, but could also include firms that implement very sophisticated agency algorithms. While most agency algorithms may not need such capabilities, and we have identified periodicity in the data that suggests lack of sophistication on the part of agency algorithms, it can definitely be the case that some low-latency activity originates from firms that do not carry out proprietary trading.

The SEC in the Concept Release on Equity Market Structure (2010) refers to high-frequency traders as "professional traders acting in a proprietary capacity that engage in strategies that generate a large number of trades on a daily basis." The SEC document suggests that firms engaging in this practice use high-speed sophisticated algorithms, employ co-location services that minimize latency, hold positions for very short intervals of time, submit and cancel many orders, and attempt to end the day with a flat inventory. Publicly-available NASDAQ data do not contain records of the accounts of each trading firms and hence the duration of their holdings and their end-of-day inventory position cannot be ascertained. Our measure attempts to pick up the other attributes (high-speed sophisticated algorithms that create dynamic strategies, fast responses that necessitate co-location, and the submission and cancellation of numerous orders) to identify the activity of high-frequency traders.[18] In this sense it is a proxy that we believe would be highly correlated with their actual activity. We stress, though, that it is a proxy rather than a direct observation of their activity.

How does *RunsInProcess* compare with constructs based on the NASDAQ HFT dataset used in Brogaard (2012), Brogaard, Hendershott, and Riordan (2012), and Carrion (in this issue)? The NASDAQ HFT dataset covers 26 high-frequency trading firms and activity in 120 stocks

[17] Hendershott, Jones, and Menkveld (2011) and Boehmer, Fong, and Wu (2012) use the number of messages in the market as a proxy for overall algorithmic activity. We believe that *RunsInProcess* is a better depiction of low-latency activity, which is a particular subset of algorithmic trading, than a simple message count.

[18] We thank a referee for this interpretation of our measure.

during 2008 and 2009. Because the identifications used to classify the high-frequency trading firms included in the dataset are not publicly available, it is difficult to validate or extend the sample. *RunsInProcess*, in contrast, is constructed from message data that are widely available for many markets and for longer time periods. However, if there are high-frequency arbitrage strategies that utilize only marketable orders, our measure may not incorporate them, though we believe that their incidence would be highly correlated with other high-frequency activities that are captured. On the other hand, the classification in the NASDAQ HFT dataset excludes the proprietary trading desks of large sell-side firms, as well as orders that are sent to the market via direct access brokers that specialize in providing services to small high-frequency trading firms.

The observations in the NASDAQ HFT dataset are executions in which the buyer and seller are classified as HF (i.e., a high-frequency trader) or non-HF, and as active or passive. The criteria used by NASDAQ to identify an HF trader are based on knowledge of the trading and order submission styles of each firm. *RunsInProcess* is based on message sequences, and so may include some agency algorithms if these carry out strategies that require low latency, though most activity by agency algorithms should be excluded when we eliminate runs shorter than 10 messages. We construct several versions of our measure that attempt to exclude activity that could be associated with agency algorithms. In one version, for example, we exclude runs that start in the first 150 ms of each second. The rationale is that the evidence presented in Section 3 could suggest that many agency algorithms operate in a periodic way and that if we exclude this period we may reduce their impact on our measure. In another version of the measure, we exclude runs where the average duration between a cancellation and a resubmission is more than 5 ms. The rationale is that high-frequency traders, but not necessarily agency algorithms, would invest in technology that would give them such capabilities. In fact, the evidence in Section 3 seems to suggest that the technology utilized by some agency algorithms is much slower. Our results using these alternative versions of our measure corroborate the conclusions we obtain from the empirical work discussed in Section 5.[19]

Since our second sample period (June 2008) overlaps with the trading data in the NASDAQ HFT dataset, we can examine the correlations between our measure of low-latency activity and four measures constructed from the HFT dataset:

1. Executed HF orders.
2. Executions with any HF participation (on either or both sides).
3. Executions against passive HF orders.
4. Executions of active non-HF against passive HF orders.

The first two measures characterize overall HF activity; the second two focus on HF executions that supply liquidity.

Of the 120 firms in the HFT dataset, 60 are NASDAQ stocks for which we use ITCH order-level data to construct the *RunsInProcess* measure.[20] Table 4 shows Spearman and Pearson correlations between the HFT-dataset measures and *RunsInProcess* over all 10-minute intervals for all stocks. As expected, *RunsInProcess* is highly correlated with high-frequency liquidity-supplying activity (the third and fourth measures). Importantly, though, *RunsInProcess* is

[19]We thank the editor, Tarun Chordia, for suggesting these ideas. Tables with the tests using these alternative versions of our measure are available from the authors.

[20]Out of the 60 stocks, 33 were in our June 2008 sample. We created the measure *RunsInProcess* for the 27 additional stocks to be able to estimate the correlations in Table 4 using all 60 stocks that are available in the HFT dataset.

Table 4
RunsInProcess and measures derived from the NASDAQ HFT dataset.

$RunsInProcess_{i,t}$ is the time-weighted average of the number of strategic runs of 10 messages or more for stock i in 10-minute interval t. The NASDAQ HFT dataset identifies the active and passive participants of trades as either high-frequency or non-high-frequency participants. For the 60 stocks common to the HFT dataset and our sample, during June, 2008, we compute, for each stock and 10-minute interval, alternative measures of high-frequency participation. We then estimate Pearson and Spearman correlations between each measure and $RunsInProcess_{i,t}$ over all 10-minute intervals for all stocks ($60 \times 819 = 49{,}410$ observations). All p-values, computed against the null hypothesis of zero correlation, are below 0.001.

		Correlation with *RunsInProcess*	
		Spearman	Pearson
Executed HF orders	Shares	0.812	0.654
	Frequency count	0.809	0.658
Executions with any HF participation (active and/or passive)	Shares	0.818	0.666
	Frequency count	0.814	0.644
Executions against passive HF orders	Shares	0.817	0.682
	Frequency count	0.810	0.634
Executions of active non-HF against passive HF orders	Shares	0.816	0.685
	Frequency count	0.809	0.643

also highly correlated with the first and second measures, which are based on total HFT trading. The Spearman correlation is over 0.8 for both the order number and share volume measures irrespective of whether these are liquidity-supplying trades or total HFT trading. Thus, our measure of low-latency activity is not restricted to solely capturing liquidity-supplying trades despite being comprised mostly of limit orders. This observation also suggests strong commonality between the liquidity-supplying and liquidity-demanding activities of high-frequency traders.

We emphasize that both our *RunsInProcess* measure and the trading measures from the HFT dataset are only proxies for the activity of high-frequency trading firms. In particular, most of the activity by high-frequency traders involves orders that do not execute. The measures computed from the HFT dataset use only executed orders, and therefore do not necessarily reflect overall activity.[21] Still, the fact that our *RunsInProcess* measure and the measures of executed orders from the HFT dataset are highly correlated should be reassuring to researchers who carry out empirical analysis using either the publicly-available ITCH data or the HFT dataset to discern the overall impact of high-frequency trading firms.

In addition to our measure of low-latency activity, we use the ITCH order-level data to compute several measures that represent different aspects of NASDAQ market quality: three measures of liquidity and a measure of short-term volatility. The first measure, *Spread*, is the time-weighted average quoted spread (ask price minus the bid price) on the NASDAQ system in an interval. The second measure, *EffSprd*, is the average effective spread (or total price impact) of all trades on NASDAQ during the 10-minute interval, where the effective spread is defined as the transaction price (quote midpoint) minus the quote midpoint (transaction price) for buy (sell) marketable orders. The

[21]The HFT dataset contains additional information, depth snapshots and quotes, for several short periods, but none of them overlaps with our sample period. Hence, we use the available information on executed orders to construct the measures we correlate with *RunsInProcess*.

third measure, *NearDepth*, is the time-weighted average number of (visible) shares in the book up to 10 cents from the best posted prices.[22] The short-term volatility measure, *HighLow*, is defined as the highest midquote in an interval minus the lowest midquote in the same interval, divided by the midpoint between the high and the low (and multiplied by 10,000 to express it in basis points).

5. Low-latency trading and market quality: analysis

To facilitate aggregation and presentation, we standardize each series at the individual stock level to have zero mean and unit variance. We first examine correlations between these standardized series. *RunsInProcess* is negatively correlated with the quoted spread (−0.32 in 2007 and −0.37 in 2008), negatively correlated with the total price impact of trades (−0.16 and −0.11 in the two sample periods, respectively), positively correlated with depth in the NASDAQ limit order book (0.29 and 0.35), and negatively correlated with short-term volatility (−0.15 and −0.24). Since all measures except depth are negative proxies for market quality, these estimates uniformly indicate a positive association between *RunsInProcess* and market quality.

We next examine the sensitivity of this association to the presence of control variables in three alternative linear models. The first model is:

$$MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} + e_{i,t}, \tag{1}$$

where $i = 1, ..., N$ indexes firms, $t = 1, ..., T$ indexes 10-minute time intervals, *MktQuality* represents one of the market quality measures (*Spread*, *EffSprd*, *NearDepth* or *HighLow*), and $TradingIntensity_{i,t}$ is stock i's total trading volume in the entire market (not just NASDAQ) in the previous 10 minute (i.e., over interval t–1). The last variable is intended to capture the impact of intraday informational events or liquidity shocks, and by construction it is predetermined. The coefficients are pooled across stocks, and the zero-mean standardization eliminates the intercept in the model.

To allow for the possibility that market-wide return and volatility factors might drive both market quality and low-latency activity, the second model is:

$$\begin{aligned} MktQuality_{i,t} = {} & a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} \\ & + a_3 R_{QQQQ,t} + a_4 |R_{QQQQ,t}| + e_{i,t}, \end{aligned} \tag{2}$$

where $R_{QQQQ,t}$ is the interval return on the NASDAQ 100 exchange traded fund, and its absolute value is a measure of volatility. We use the NASDAQ index in preference to the S&P index because NASDAQ is the primary listing exchange for all stocks in our sample.

Our third specification is motivated by theoretical models that give rise to intraday patterns in liquidity (as well as various empirical findings of time-of-day effects in liquidity measures). For example, models of adverse selection (e.g., Glosten and Milgrom, 1985) generally predict higher spreads in the morning compared to the rest of the day. An afternoon increase in spreads is consistent with inelasticity of demand (e.g., Brock and Kleidon, 1992), while the analysis in Admati and Pfleiderer (1988) could be used to justify morning and afternoon patterns driven by implicit or explicit coordination of traders in the market. Accordingly, the third model

[22] We stress that spreads, effective spreads (or total price impact), and depth are measures of liquidity rather than direct measures of investor transaction costs. Investors can use both marketable orders (that pay the spread) and limit orders (that earn the spread), and hence the cost of execution for a certain position depends on whether investors utilize dynamic strategies consisting of both limit and marketable orders. While such dynamic strategies would no doubt incorporate estimates of the market quality measures we study, the relationship between the cost arising from the optimal strategy and these measures may not be monotonic.

incorporates morning and afternoon dummy variables:

$$MktQuality_{i,t} = a_0 + a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} + a_3 DumAM_t + a_4 DumPM_t + e_{i,t}, \quad (3)$$

where $DumAM_t$ is equal to one for intervals between 9:30 am and 11:00 am, and zero otherwise, and $DumPM_t$ is equal to one for intervals between 2:30 pm and 4:00 pm, and zero otherwise. Despite the zero-mean standardization, an intercept is required in this specification to capture the residual mid-day effects (11:00 am–2:30 pm).

Table 5 reports estimates of models (1), (2) and (3). The coefficient estimates are OLS, and the standard errors are computed using the Driscoll-Kraay extension of the Newey-West HAC estimator (Driscoll and Kraay, 1998; Baum, Schaffer, and Stillman, 2010; Thompson, 2011). The Discoll-Kray procedure is a GMM technique for panels where both the cross-sectional and time dimensions are large. The moment conditions are constructed as cross-sectional averages, and so the estimates effectively cluster on time. The GMM framework is useful in that it readily accommodates the Newey-West autocorrelation correction. The coefficient estimates are identical to the OLS estimates, but the standard errors are in principle robust to heteroscedasticity, autocorrelation, and general spatial (cross-firm) dependence.

Panel A of Table 5 presents estimated coefficients for model (1) in the 2007 and 2008 samples. The most interesting coefficient is a_1, which measures the association of low-latency activity with the market quality measures. We observe that higher low-latency activity is associated with lower posted and effective spreads, greater depth, and lower short-term volatility. Moreover, the relationship between low-latency activity and market quality is similar in the 2007 and 2008 sample periods. Estimates for models (2) and (3) (in Panels B and C) are generally similar.

The OLS estimates cannot be interpreted as causal impact coefficients due to the possibility of endogeneity arising from simultaneity. For example, an exogenous drop in spreads might establish a more attractive environment for, and lead to an increase in, low-latency activity. This mechanism would induce correlation between *RunsInProcess* and the errors of the regression, rendering OLS estimates inconsistent.

There are no clear and obvious candidates for instrumental variables. Most constructs based on stock-specific data for a given interval are so closely related to *RunsInProcess* or the market quality measures that they are subject to the same simultaneity concerns. We nevertheless believe that an effort to separate the impact of low-latency activity on spread, depth, and volatility from influences working in the opposite direction is warranted. We therefore turn to construction of instruments and specifications that might resolve, albeit imperfectly, the two effects.

A good instrument for low-latency activity in, say, model (1), should satisfy two requirements. It should be correlated with $RunsInProcess_{i,t}$ and it should also be uncorrelated with the $e_{i,t}$ disturbance. If low-latency activity has a significant market-wide component, then a market-wide average of *RunsInProcess* is likely to satisfy the first requirement. Market-wide factors in low-latency activity could plausibly arise from funding constraints or inventory risk management at the high-frequency trading firms. The experience of the flash crash documented in Kirilenko, Kyle, Samadi, and Tuzun (2011) and discussed in the joint CFTC/SEC report is one illustration (albeit dramatic) of this tendency (U.S. Commodities Futures Trading Commission and the U.S. Securities and Exchange Commission, 2010).

The second requirement (absence of correlation with $e_{i,t}$) is more difficult to achieve. We can eliminate one obvious source of correlation, though, by excluding from the broader *RunsInProcess* average the contributions from $RunsInProcess_{i,t}$ (the variable we are attempting to instrument) and also contributions from stocks likely to be closely related to stock i by reason

Table 5
Low-latency trading and market quality: OLS estimates.

This table presents pooled panel regression analyses that relate low-latency activity to market quality. The low-latency activity measure is $RunsInProcess_{i,t}$, the time-weighted average of the number of strategic runs of 10 messages or more for stock i in10-minute interval t. $MktQuality_{i,t}$ is a placeholder denoting: $Spread_{i,t}$, the quoted spread; $EffSprd_{i,t}$, the effective spread; $NearDepth_{i,t}$, book depth within 10 cents of the best bid or offer; or, $HighLow_{i,t}$, the midquote range divided by the midquote average. $TradingIntensity_{i,t}$ is the stock's total trading volume during the previous 10 minutes. $R_{QQQQ,t}$is the return on the NASDAQ 100 index ETF. $DumAM_t$ is a morning dummy variable (equal to one between 9:30 am and 11:00 am); $DumPM_t$ is an afternoon dummy variable (2:30 pm to 4:00 pm). We standardize each (non-dummy) variable by subtracting from each observation the stock-specific time-series average and dividing by the stock-specific time-series standard deviation. Coefficient estimates are OLS; p-values (against a two-sided alternative) are computed using the Driscoll-Kraay extension of the Newey-West estimator.

Panel A. Model (1): $MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} + e_{i,t}$

		2007		2008	
		a_1	a_2	a_1	a_2
Spread	Coeff. (p-value)	−0.214 (<0.001)	0.096 (<0.001)	−0.275 (<0.001)	0.046 (0.036)
EffSprd	Coeff. (p-value)	−0.170 (<0.001)	0.103 (<0.001)	−0.193 (<0.001)	0.038 (0.001)
NearDepth	Coeff. (p-value)	0.246 (<0.001)	−0.104 (<0.001)	0.297 (<0.001)	−0.003 (0.888)
HighLow	Coeff. (p-value)	−0.056 (<0.001)	0.322 (<0.001)	−0.117 (<0.001)	0.229 (<0.001)

Panel B. Model (2): $MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} + a_3 R_{QQQ,t} + a_4 |R_{QQQQ,t}| + e_{i,t}$

		2007				2008			
		a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
Spread	Coeff. (p-value)	−0.204 (<0.001)	0.080 (<0.001)	−0.010 (0.654)	0.171 (<0.001)	−0.267 (<0.001)	0.039 (0.093)	−0.014 (0.533)	0.141 (<0.001)
EffSprd	Coeff. (p-value)	−0.164 (<0.001)	0.094 (<0.001)	−0.005 (0.680)	0.095 (<0.001)	−0.189 (<0.001)	0.034 (0.003)	−0.006 (0.687)	0.076 (<0.001)
NearDepth	Coeff. (p-value)	0.238 (<0.001)	−0.092 (<0.001)	0.031 (0.086)	−0.131 (<0.001)	0.292 (<0.001)	0.001 (0.977)	−0.014 (0.539)	−0.081 (<0.001)
HighLow	Coeff. (p-value)	−0.038 (0.002)	0.295 (<0.001)	−0.007 (0.678)	0.298 (<0.001)	−0.099 (<0.001)	0.215 (<0.001)	0.003 (0.890)	0.309 (<0.001)

Panel C. Model (3): $MktQuality_{i,t} = a_0 + a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} + a_3 DumAM_t + a_4 DumPM_t + e_{i,t}$

		2007					2008				
		a_0	a_1	a_2	a_3	a_4	a_0	a_1	a_2	a_3	a_4
Spread	Coeff. (p-value)	−0.089 (<0.001)	−0.151 (<0.001)	0.069 (<0.001)	0.678 (<0.001)	−0.224 (<0.001)	−0.142 (<0.001)	−0.144 (<0.001)	0.011 (0.546)	0.931 (<0.001)	−0.226 (<0.001)
EffSprd	Coeff. (p-value)	−0.060 (<0.001)	−0.129 (<0.001)	0.085 (<0.001)	0.437 (<0.001)	−0.134 (<0.001)	−0.082 (<0.001)	−0.115 (<0.001)	0.018 (0.053)	0.548 (<0.001)	−0.143 (<0.001)
NearDepth	Coeff. (p-value)	0.018 (0.415)	0.191 (<0.001)	−0.090 (<0.001)	−0.538 (<0.001)	0.402 (<0.001)	0.031 (0.036)	0.165 (<0.001)	0.004 (0.777)	−0.719 (<0.001)	0.509 (<0.001)
HighLow	Coeff. (p-value)	−0.152 (<0.001)	−0.004 (0.775)	0.286 (<0.001)	0.617 (<0.001)	0.095 (0.034)	−0.244 (<0.001)	−0.009 (0.346)	0.170 (<0.001)	1.004 (<0.001)	0.141 (<0.001)

of common trading strategies. By way of illustration, consider the error term in Model I applied to the *Spread* measure of market quality. The time-t error for, say, Apple ($e_{AAPL,t}$) inherently affects $Spread_{AAPL,t}$, which may in turn affect $RunsInProcess_{AAPL,t}$. However, $e_{AAPL,t}$ would not directly affect *RunsInProcess* for other stocks, with the possible exception of other computer stocks and other NASDAQ 100 stocks or S&P 500 stocks. The latter connections might arise from the pursuit of index strategies and within-industry pairs strategies.

The instrument for $RunsInProcess_{i,t}$ that we propose, denoted $RunsNotIND_{i,t}$, is the average number of runs of 10 messages or more across all the stocks in our sample excluding: (1) the INDividual stock, stock i, (2) stocks in the same INDustry as stock i (as defined by the four-digit SIC code), and (3) stocks in the same INDex as stock i, if stock i belongs to either the NASDAQ 100 Index or the S&P 500 Index. Our hope is that by excluding the most likely candidates for such cross-stock strategies, $RunsNotIND_{i,t}$ would not be affected by the liquidity and volatility of stock i, strengthening the economic rationale for using it as an instrument.

Beyond stocks in the same industry and the same index that are explicitly omitted from the instrument, our results should be robust to multi-stock algorithms that utilize concurrent trading in a small number of stocks. The average (minimum) number of stocks that are used in the constructions of *RunsNotIND* is 322.7 (250) in 2007 and 371.3 (290) in 2008, making it insensitive to concurrent trading in a handful of related stocks. For robustness, we repeated the analysis with an instrument computed as the median of $RunsInProcess_{i,t}$ (excluding stock i, stocks in the same industry, and stocks in the same index) in each interval because the median should be even less sensitive to outliers. Our results with the median instrument are similar to those with *RunsNotIND*, suggesting that multi-stock algorithms are not a significant problem with respect to the validity of this instrument.[23]

The exclusions used in the construction of *RunsNotIND* are motivated by the second requirement for a valid instrument, but they do not impair its ability to satisfy the first requirement: the correlation between $RunsInProcess_{i,t}$ and $RunsNotIND_{i,t}$ (pooled across all stocks and time intervals) is 0.521. We therefore estimate models (1), (2) and (3) with $RunsNotIND_{i,t}$ as an instrument for $RunsInProcess_{i,t}$. The reported coefficient estimates are 2SLS, and the standard errors are (as above) Driscoll-Kraay with two-way clustering.

Panel A of Table 6 presents the estimated coefficients of model (1) side-by-side for the 2007 and 2008 sample periods. Contingent on the validity of our instrument, the coefficient a_1 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. In fact, the results appear somewhat stronger than those using the simple OLS.[24] In all regressions, Cragg-Donald (1993) statistics reject the null of weak instruments using the Stock and Yogo (2005) critical values.

Panel B of Table 6 suggests that the influence of low-latency activity is not driven by omitted variables related to market return or volatility. In fact, the market's return is almost

[23]If some market participants implement complex algorithms whereby a single co-located algorithm makes low-latency trading decisions in a large number of stocks (say hundreds of stocks), the quality of our instrument may suffer. In fact, extensive trading of this nature would invalidate the instrument. However, talking to industry participants and regulators led us to believe that a single algorithm (i.e., a distinct process that runs on a co-located machine) is usually not responsible for high-frequency trading in many securities. Rather, each algorithm implements a strategy in a more limited set of securities, often a single stock. In particular, the speed requirements of competitive low-latency trading give rise to simpler, rather than more complex, algorithms.

[24]The finding that the results strengthen when the 2SLS instrument effectively removes noise from each stock's $RunsInProcess_{i,t}$ could suggest a prominent role for market-wide determinants of low-latency activity.

Table 6
Low-latency trading and market quality: 2SLS estimates.

The table presents pooled panel regression analyses that relate low-latency activity to market quality. The low-latency activity measure is $RunsInProcess_{i,t}$, the time-weighted average of the number of strategic runs of 10 messages or more for stock i in 10-minute interval t. As an instrument for $RunsInProcess_{i,t}$ we use $RunsNotIND_{i,t}$, which is the average of *RunsInProcess* for all other stocks, excluding stock i itself, stocks in the same four-digit SIC industry as i, and stocks in the same index as i. $MktQuality_{i,t}$ is a placeholder denoting: $Spread_{i,t}$, the quoted spread; $EffSprd_{i,t}$, the effective spread; $NearDepth_{i,t}$, book depth within 10 cents of the best bid or offer; and, $HighLow_{i,t}$, the midquote range divided by the midquote average. $TradingIntensity_{i,t}$ is the stock's total trading volume during the previous 10 minutes. $R_{QQQQ,t}$ is the return on the NASDAQ 100 index ETF. $DumAM_t$ is a morning dummy variable (equal to one between 9:30 am and 11:00 am); $DumPM_t$ is an afternoon dummy variable (2:30 pm to 4:00 pm). We standardize each (non-dummy) variable by subtracting from each observation the stock-specific time-series average and dividing by the stock-specific time-series standard deviation. Coefficient estimates are 2SLS; p-values (against a two-sided alternative) are computed using the Driscoll-Kraay extension of the Newey-West estimator.

Panel A. Model (1): $MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} + e_{i,t}$

		2007		2008	
		a_1	a_2	a_1	a_2
Spread	Coeff. (p-value)	–0.682 (<0.001)	0.104 (<0.001)	–0.959 (<0.001)	0.050 (<0.001)
EffSprd	Coeff. (p-value)	–0.493 (<0.001)	0.109 (<0.001)	–0.654 (<0.001)	0.041 (<0.001)
NearDepth	Coeff. (p-value)	0.460 (<0.001)	–0.107 (<0.001)	0.804 (<0.001)	–0.007 (0.729)
HighLow	Coeff. (p-value)	–0.437 (<0.001)	0.329 (<0.001)	–0.648 (<0.001)	0.233 (<0.001)

Panel B. Model (2): $MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} + a_3 R_{QQQQ,t} + a_4 |R_{QQQQ,t}| + e_{i,t}$

		2007				2008			
		a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
Spread	Coeff. (p-value)	–0.642 (<0.001)	0.090 (<0.001)	–0.007 (0.719)	0.146 (<0.001)	–0.939 (<0.001)	0.045 (0.001)	–0.011 (0.481)	0.102 (<0.001)
EffSprd	Coeff. (p-value)	–0.471 (<0.001)	0.101 (<0.001)	–0.004 (0.759)	0.077 (<0.001)	–0.644 (<0.001)	0.039 (<0.001)	–0.004 (0.682)	0.050 (<0.001)
NearDepth	Coeff. (p-value)	0.426 (<0.001)	–0.096 (<0.001)	0.030 (0.093)	–0.120 (<0.001)	0.794 (<0.001)	–0.004 (0.825)	–0.015 (0.438)	–0.052 (0.004)
HighLow	Coeff. (p-value)	–0.360 (<0.001)	0.302 (<0.001)	–0.005 (0.752)	0.280 (<0.001)	–0.593 (<0.001)	0.220 (<0.001)	0.004 (0.774)	0.281 (<0.001)

Panel C. Model (3): $MktQuality_{i,t} = a_0 + a_1 RunsInProcess_{i,t} + a_2 TradingIntensity_{i,t} + a_3 DumAM_t + a_4 DumPM_t + e_{i,t}$

		2007					2008				
		a_0	a_1	a_2	a_3	a_4	a_0	a_1	a_2	a_3	a_4
Spread	Coeff. (p-value)	–0.051 (0.055)	–0.448 (<0.001)	0.083 (<0.001)	0.514 (<0.001)	–0.240 (<0.001)	–0.120 (<0.001)	–0.552 (<0.001)	0.019 (0.198)	0.676 (<0.001)	–0.094 (0.027)
EffSprd	Coeff. (p-value)	–0.032 (0.054)	–0.349 (<0.001)	0.095 (<0.001)	0.317 (<0.001)	–0.145 (<0.001)	–0.063 (<0.001)	–0.462 (<0.001)	0.025 (<0.001)	0.331 (<0.001)	–0.030 (0.218)
NearDepth	Coeff. (p-value)	0.017 (0.471)	0.199 (<0.001)	–0.091 (<0.001)	–0.534 (<0.001)	0.402 (<0.001)	0.031 (<0.001)	0.155 (0.051)	0.004 (0.781)	–0.725 (<0.001)	0.512 (<0.001)
HighLow	Coeff. (p-value)	–0.120 (<0.001)	–0.255 (<0.001)	0.298 (<0.001)	0.480 (<0.001)	0.082 (0.068)	–0.225 (<0.001)	–0.367 (<0.001)	0.177 (<0.001)	0.781 (<0.001)	0.257 (<0.001)

never statistically significant. While market volatility is a significant driver behind all of our measures of market quality, its inclusion has a negligible effect on the a_1 coefficients when we examine spreads, effective spreads, and depth, and only a minor effect when the dependent variable is short-term volatility. Panel C of Table 6 shows the results with time-of-day dummies. In general, the time-dummy coefficients suggest that the market is less liquid in the first hour and a half of trading and more liquid in the afternoon. However, time-of-day effects do not eliminate the impact of low-latency activity, testifying to the robustness of the effects we document.

This analysis extends our understanding of the impact of low-latency activity on market quality by reducing the likelihood that the results are driven by reverse causality (i.e., the impact of the liquidity or volatility of a particular stock in the particular interval on $RunsInProcess_{i,t}$). We note that despite the possible limitations of our instrument, the fact that the results strengthen when we move from the OLS to the instrumental variables estimation is consistent with low-latency trading impacting market quality.

The next logical step is joint modeling of both the market quality measures and low-latency activity in a simultaneous-equations framework. While this would necessitate use of a separate instrument for the market quality measures, it offers at least one significant advantage. In models (1)-(3) we use lagged volume to capture stock-specific informational events or liquidity shocks. With a second equation, we could use a contemporaneous instrument in lieu of the lagged variable, and hence have a better control for stock-specific conditions in the same interval over which we measure the low-latency activity.

The instrument we use for market quality in the simultaneous-equations specifications is $EffSprdNotNAS_{i,t}$, which 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 but only from trades executed on non-NASDAQ trading venues (using the TAQ database). This measure 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, rendering it an appropriate instrument.

It might be argued that $EffSprdNotNAS_{i,t}$ would not be exogenous if many low-latency algorithms pursue cross-market strategies in the same security (e.g., if a single algorithm co-located with NASDAQ 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.[25] 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 to the different markets. 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 ($RunsInProcess_{i,t}$) and hence would not introduce reverse causality.

We emphasize that the goal of this analysis is simply investigation of alternative approaches to addressing endogeneity concerns. The models are not inherently superior to the single-equation models, but their estimation is contingent on a modified set of assumptions, replacing lagged

[25]Conversations with a NASDAQ official in 2010 provided support to this view.

volume on the right-hand side with an attempt to explicitly model the contemporaneous market quality measures.

We specify three two-equation econometric models, which are derived from the first set (models (1)-(3)). Model (4) extends model (1) as:

$$\begin{aligned} MktQuality_{i,t} &= a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} + e_{1,i,t} \\ RunsInProccess_{i,t} &= b_1 MktQuality_{i,t} + b_2 RunsNotIND_{i,t} + e_{2,i,t}. \end{aligned} \tag{4}$$

The instruments are $EffSprdNotNAS_{i,t}$ and $RunsNotIND_{i,t}$. Model (5) extends model (2) to admit common return and volatility measures:

$$\begin{aligned} MktQuality_{i,t} &= a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} \\ &\quad + a_3 R_{QQQQ,t} + a_4 |R_{QQQQ,t}| + e_{1,i,t} \\ RunsInProccess_{i,t} &= b_1 MktQuality_{i,t} + b_2 RunsNotI_{i,t} \\ &\quad + a_3 R_{QQQQ,t} + a_4 |R_{QQQQ,t}| + e_{2,i,t}. \end{aligned} \tag{5}$$

Model (6) introduces time-of-day dummy variables, following model (3):

$$\begin{aligned} MktQuality_{i,t} &= a_0 + a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} \\ &\quad + a_3 DumAM_t + a_4 DumPM_t + e_{1,i,t} \\ RunsInProccess_{i,t} &= b_0 + b_1 MktQuality_{i,t} + b_2 RunsNotI_{i,t} \\ &\quad + b_3 DumAM_t + b_4 DumPM_t + e_{2,i,t}. \end{aligned} \tag{6}$$

As in the earlier models, we estimate the coefficients in these systems using 2SLS, and report robust standard errors (Driscoll-Kraay with two-way clustering). Panel A of Table 7 presents the estimated coefficients of Model IV side-by-side for the 2007 and 2008 sample periods. As before, we observe that higher low-latency activity implies lower posted and effective spreads, greater depth, and lower short-term volatility. The results in the presence of market factors (model (5), Panel B) and time-of-day dummy variables (model 6, Panel C) generally agree. In all regressions, Cragg-Donald (1993) statistics reject the null of weak instruments using the Stock and Yogo (2005) critical values.

To gauge the economic magnitudes implied by the a_1 coefficients, we compute the change in a market quality measure for a representative stock implied by a given increase in low-latency activity. A one standard deviation increase in *RunsInProcess* implies a decrease of 26% in spreads in the 2007 sample period and a decrease of 32% in the 2008 sample period. A similar pattern whereby low-latency activity has a greater positive impact on market quality in 2008 is also observed for depth within 10 cents from the best prices, where one standard deviation increase in *RunsInProcess* implies an increase by 20% in the 2007 sample period (up 2,199 shares from a mean of 11,271 shares) and an even greater increase (34%) is observed in the 2008 sample period when the market is under stress. A one standard deviation increase in *RunsInProcess* also implies a decrease in short-term volatility of 29% in 2007 (down 12.3 basis points from a mean value of 42.1 basis points) and 32% in 2008.

The fact that low-latency activity lowers spreads, increases depth, and decreases short-term volatility even to a greater extent in the 2008 sample period—when 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. It could be that greater variability in the measures during stressful times simply means that statistical methods are better able to identify the relationships. However, there could be economic reasons for why we might observe this

Table 7
Low-latency trading and market quality: simultaneous equation estimates.

The table presents pooled panel regression analyses that relate low-latency activity to market quality. The low-latency activity measure is $RunsInProcess_{i,t}$, the time-weighted average of the number of strategic runs of 10 messages or more for stock i in 10-minute interval t. As an instrument for $RunsInProcess_{i,t}$ we use $RunsNotIND_{i,t}$, which is the average of *RunsInProcess* for all other stocks, excluding stock i itself, stocks in the same four-digit SIC industry as i, and stocks in the same index as i. $MktQuality_{i,t}$ is a placeholder denoting: $Spread_{i,t}$, the quoted spread; $EffSprd_{i,t}$, the effective spread; $NearDepth_{i,t}$, book depth within 10 cents of the best bid or offer; and, $HighLow_{i,t}$, the midquote range divided by the midquote average. As an instrument for $MktQuality_{i,t}$ we use $EffSprdNotNas_{i,t}$, which is the average dollar effective spread from executions on other (non-NASDAQ) trading venues. $R_{QQQQ,t}$ is the return on the NASDAQ 100 index ETF. $DumAM_t$ is a morning dummy variable (equal to one between 9:30 am and 11:00 am); $DumPM_t$ is an afternoon dummy variable (2:30 pm to 4:00 pm). We standardize each (non-dummy) variable by subtracting from each observation the stock-specific time-series average and dividing by the stock-specific time-series standard deviation. Coefficient estimates are 2SLS; p-values (against a two-sided alternative) are computed using the Driscoll-Kraay extension of the Newey-West estimator.

Panel A. Model (4):

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

		2007				2008			
		a_1	a_2	b_1	b_2	a_1	a_2	b_1	b_2
Spread	Coeff.	–0.534	0.567	–0.052	0.494	–0.615	0.526	–0.107	0.461
	(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
EffSprd	Coeff.	–0.203	0.382	–0.079	0.500	–0.143	0.219	–0.265	0.475
	(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
NearDepth	Coeff.	0.380	–0.237	0.123	0.484	0.716	–0.116	0.379	0.359
	(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
HighLow	Coeff.	–0.350	0.476	–0.063	0.497	–0.475	0.452	–0.126	0.464
	(p-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)

Panel B. Model (5):

$$MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} + a_3 R_{QQQQ,t} + a_4 |R_{QQQQ,t}| + e_{1,i,t}$$
$$RunsInProccess_{i,t} = b_1 MktQuality_{i,t} + b_2 RunsNotIND_{i,t} + a_3 R_{QQQQ,t} + a_4 |R_{QQQQ,t}| + e_{2,i,t}$$

			a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
2007	*Spread*	Coeff.	–0.471	0.539	–0.001	0.139	–0.059	0.496	0.001	0.017
		(p-value)	(<0.001)	(<0.001)	(0.963)	(<0.001)	(<0.001)	(<0.001)	(0.878)	(0.004)
	EffSprd	Coeff.	–0.167	0.365	0.051	0.074	–0.089	0.502	0.005	0.015
		(p-value)	(<0.001)	(<0.001)	(0.012)	(0.002)	(<0.001)	(<0.001)	(0.179)	(0.011)
	NearDepth	Coeff.	0.341	–0.221	0.044	–0.095	0.141	0.485	–0.006	0.022
		(p-value)	(<0.001)	(<0.001)	(0.005)	(<0.001)	(<0.001)	(<0.001)	(0.192)	(0.001)
	HighLow	Coeff.	–0.240	0.428	–0.032	0.251	–0.075	0.501	–0.002	0.028
		(p-value)	(<0.001)	(<0.001)	(0.027)	(<0.001)	(<0.001)	(<0.001)	(0.609)	(<0.001)

Table 7 (*continued*)

Panel B. Model (5):

$$MktQuality_{i,t} = a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} + a_3 R_{QQQQ,t} + a_4 |R_{QQQQ,t}| + e_{1,i,t}$$
$$RunsInProccess_{i,t} = b_1 MktQuality_{i,t} + b_2 RunsNotIND_{i,t} + a_3 R_{QQQQ,t} + a_4 |R_{QQQQ,t}| + e_{2,i,t}$$

			a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
2008	*Spread*	Coeff.	–0.595	0.509	0.040	0.124	–0.105	0.461	–0.011	–0.013
		(*p*-value)	(<0.001)	(<0.001)	(0.254)	(0.001)	(<0.001)	(<0.001)	(0.219)	(0.147)
	EffSprd	Coeff.	–0.132	0.210	0.022	0.066	–0.263	0.475	–0.010	–0.009
		(*p*-value)	(<0.001)	(<0.001)	(0.291)	(<0.001)	(<0.001)	(<0.001)	(0.361)	(0.421)
	NearDepth	Coeff.	0.713	–0.114	–0.056	–0.047	0.369	0.362	0.009	–0.003
		(*p*-value)	(<0.001)	(<0.001)	(0.040)	(0.064)	(<0.001)	(<0.001)	(0.542)	(0.835)
	HighLow	Coeff.	–0.448	0.430	0.176	0.243	–0.126	0.464	0.007	0.005
		(*p*-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.447)	(0.629)

Panel C. Model (6):

$$MktQuality_{i,t} = a_0 + a_1 RunsInProcess_{i,t} + a_2 EffSprdNotNAS_{i,t} + a_3 DumAM_t + a_4 DumPM_t + e_{1,i,t}$$
$$RunsInProccess_{i,t} = b_0 + b_1 MktQuality_{i,t} + b_2 RunsNotIND_{i,t} + b_3 DumAM_t + b_4 DumPM_t + e_{2,i,t}$$

			a_0	a_1	a_2	a_3	a_4	b_0	b_1	b_2	b_3	b_4
2007	*Spread*	Coeff.	–0.022	–0.437	0.555	0.215	–0.121	0.002	–0.055	0.498	0.016	–0.026
		(*p*-value)	(0.187)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.734)	(<0.001)	(<0.001)	(0.358)	(0.085)
	EffSprd	Coeff.	–0.002	–0.184	0.380	0.039	–0.032	0.004	–0.082	0.502	0.008	–0.023
		(*p*-value)	(0.846)	(<0.001)	(<0.001)	(0.002)	(<0.001)	(0.606)	(<0.001)	(<0.001)	(0.679)	(0.133)
	NearDepth	Coeff.	0.033	0.150	–0.207	–0.494	0.350	–0.001	0.149	0.499	0.078	–0.072
		(*p*-value)	(0.121)	(0.003)	(<0.001)	(<0.001)	(<0.001)	(0.879)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
	HighLow	Coeff.	–0.127	–0.220	0.466	0.372	0.183	–0.005	–0.067	0.503	0.029	–0.008
		(*p*-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.508)	(<0.001)	(<0.001)	(0.093)	(0.630)
2008	*Spread*	Coeff.	–0.051	–0.566	0.518	0.160	0.063	0.033	–0.095	0.475	–0.049	–0.094
		(*p*-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.094)	(0.001)	(<0.001)	(<0.001)	(0.022)	(0.003)
	EffSprd	Coeff.	–0.006	–0.143	0.218	0.011	0.013	0.037	–0.232	0.485	–0.062	–0.099
		(*p*-value)	(0.330)	(<0.001)	(<0.001)	(0.376)	(0.319)	(<0.001)	(<0.001)	(<0.001)	(0.005)	(0.002)
	NearDepth	Coeff.	0.037	0.160	–0.137	–0.646	0.486	0.025	0.361	0.473	0.169	–0.275
		(*p*-value)	(0.011)	(0.023)	(<0.001)	(<0.001)	(<0.001)	(0.013)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
	HighLow	Coeff.	–0.194	–0.383	0.417	0.508	0.335	0.015	–0.120	0.479	–0.004	–0.061
		(*p*-value)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.130)	(<0.001)	(<0.001)	(0.857)	(0.046)

effect. It is reasonable to assume that higher volatility creates more profit opportunities for high-frequency traders. Even smaller stocks that normally are not very attractive to high-frequency traders due to the lack of volume can become profitable enough to warrant their attention during those times.

We find evidence consistent with this intuition when estimating the models separately on four quartiles ranked by the average market capitalization over the sample period. The a_1 coefficients in the subsamples have the same sign as in the full sample, and are all statistically significant. While there is no consistent pattern across the quartiles in the 2007 sample period, 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. Hence, it is conceivable that the stronger results in the 2008 sample period are at least partially driven by increased activity of high-frequency traders in smaller stocks.[26]

To investigate whether our pooled estimates might be unduly influenced by outliers, we estimate model (4) stock-by-stock. Fig. 3 presents histograms of the a_1 coefficient estimates. The first two panels of Fig. 3, for example, show that almost all of the a_1 coefficients are negative when the market quality measure is the quoted spread. The histograms of all other market quality measures demonstrate that the pooled results are not driven by outliers but rather represent a reasonable summary of the manner in which low-latency activity affects market quality in the cross-section of stocks.

Panel B of Table 7 presents the results of model (5), where we add common factor information (return and volatility of the market) to the simultaneous-equations model. Market volatility appears to be an important determinant of the market quality measures in both sample periods (the a_4 coefficient). As a determinant of low-latency activity, market volatility is significant only in 2007 (the b_4 coefficient). Market return has an impact on some of the market quality measures (especially depth and short-term volatility), but is not significant in the *RunsInProcess* equation. The estimates in this panel suggest that the inclusion of market return and volatility as independent variables does not eliminate the significant showing of the low-latency activity as a determinant of the market quality measures: all estimated a_1 coefficients have the same signs as in Panel A of Table 7 and are highly statistically significant.

Similar results are found in the estimates of model (6) (in Panel C of Table 7), which includes dummy variables to account for potential time-of-day effects. As in Table 6, we observe that time-of-day variables exert their expected influence on market activity. The simultaneous-equation model also allows us to see that the intensity of low-latency trading is lower in the last hour and a half of trading (the b_4 coefficient). Finally, the column of a_1 coefficients clearly shows that our results that greater low-latency trading implies lower spreads and effective spreads, greater depth, and lower short-term volatility remain extremely robust.

6. Related literature

Our paper can be viewed from two related angles: (1) speed of information dissemination and activity in financial markets, and (2) high-frequency trading (or algorithmic trading in general) and its impact on the market environment.

[26]We also estimate the specifications separately on 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, however, show no significant patterns across market-share quartiles. In other words, the beneficial impact of low-latency trading on the market quality measures is similar for stocks that have varying degrees of trading concentration on the NASDAQ system. The subsample results are available from the authors upon request.

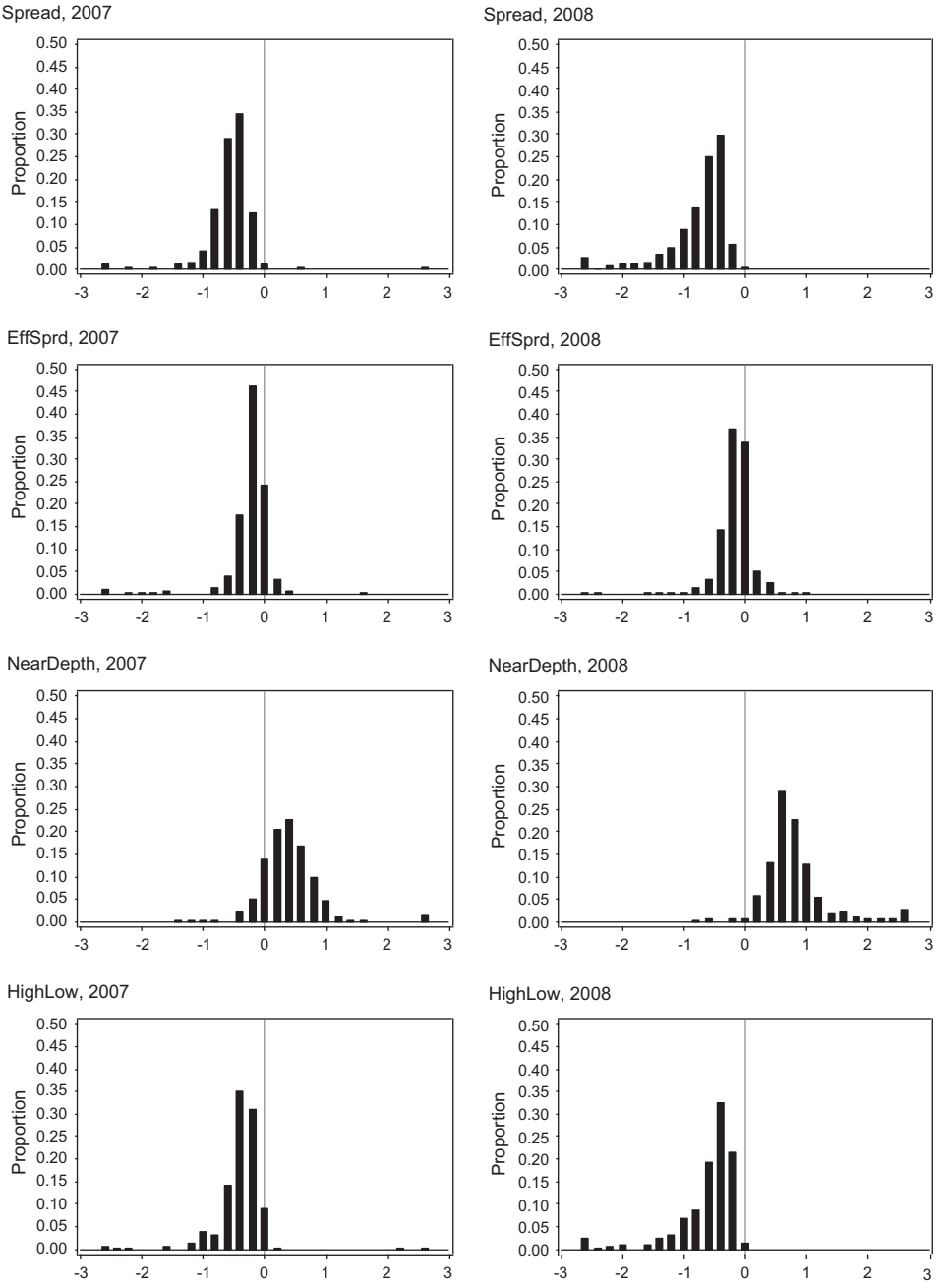

Fig. 3. Histogram of a_1 coefficients in simultaneous equation estimates at the firm level. For each of the market quality measures (*HighLow*, *EffSprd*, *Spread*, and *NearDepth*) the specification [model (4)],

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

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

is estimated at the individual firm level using 2SLS with instruments *RunsNotIND* and *EffSprdNotNas*. Each histogram depicts the distribution of the a_1 coefficient estimates for the indicated year and market quality measure.

Regarding speed, Hendershott and Moulton (2011) 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 10 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 (2012) find that a reduction in latency (from 50 to 10 ms) on the Deutsche Boerse' Xetra system was 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 crowded out traditional liquidity providers) and the sophistication of liquidity demanders (e.g., their adoption of algorithms to implement dynamic limit order strategies that can both supply and demand liquidity).[27]

Early papers on algorithmic trading sought to establish stylized facts related to algorithmic activity (Prix, Loistl, and Huetl, 2007; Gsell, 2009; Gsell and Gomber, 2009; Groth, 2009), while later research evaluated its impact on the market (Chaboud, Hjalmarsson, Vega and Chiquoine, 2013; Hendershott, Jones, and Menkveld, 2011; Boehmer, Fong, and Wu, 2012; Hendershott and Riordan, 2013).

In particular, Hendershott, Jones, and Menkveld (2011) 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 an increase in algorithmic activity produces mixed results: quoted and effective spreads go down, but so does quoted depth. 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.[28] Two considerations could account for the difference in 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 algorithms. The 2003 autoquoting change, therefore, may have mostly affected the activity of agency algorithms.

Several recent papers focus on the activity of high-frequency traders in an attempt to characterize the impact of these proprietary algorithms on the market environment. Brogaard (2012) investigates the impact of high-frequency trading on market quality using two special datasets of 120 stocks: the aforementioned HFT dataset from NASDAQ, as well as another one from BATS with 25 high-frequency traders. He reports that high-frequency traders contribute to liquidity provision and that their activity appears to lower volatility. While he uses a measure of high-frequency trading different from the one we propose, his findings are consistent with ours.

[27] Cespa and Foucault (forthcoming) and Easley, O'Hara, and Yang (2013) 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. Pagnotta and Philippon (2012) model speed as a differentiating attribute of competing exchanges. Moallemi and Sağlam (2013) 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.

[28] The average market capitalization (in billion dollars) of sample quintiles reported in Table 1 of Hendershott, Jones, and Menkveld (2011) 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.

Brogaard, Hendershott, and Riordan (2012) use the NASDAQ HFT dataset to investigate the role high-frequency trading plays in price discovery. They estimate a model of price formation and report that when high-frequency trading firms trade by demanding liquidity, they do so in the direction of the permanent price changes and in the opposite direction to transitory price changes. Hence, they conclude that high-frequency traders help price efficiency.[29]

Three other papers on high-frequency trading are published in this special issue of the *Journal of Financial Markets*. Carrion (in this issue) uses the NASDAQ HFT dataset to show that the aggregate profit patterns of high-frequency traders point to intraday market timing skill, but not necessarily the ability to profit from minute-by-minute movements. High-frequency traders also appear to supply liquidity when it is scarce and demand it when it is more plentiful. Menkveld (in this issue) looks at the entry of a large high-frequency trader that acts predominantly as a multi-venue market maker in Europe. The study characterizes the profits of the high-frequency trader and shows how its net position interacts with the price process. Hagströmer and Norden (in this issue) have special data from NASDAQ OMX Stockholm that enable them to characterize the strategies of high-frequency trading firms. They categorize high-frequency traders as either "market makers" or "opportunistic traders", and find that those pursuing market making strategies appear to mitigate volatility in the market.

While we [as well as Brogaard (2012), Brogaard, Hendershott, and Riordan (2012), and Carrion (in this issue)] find a positive impact on market quality, 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 several high-frequency traders in the equity markets 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. Similarly, Kirilenko, Kyle, Samadi, and Tuzun (2011) investigate the behavior of high-frequency trading firms in the futures market during the flash crash. They define "high-frequency traders" in the S&P 500 E-mini futures contract as those traders that execute a large number of daily transactions and fit a certain profile of intraday and end-of-day net positions. The authors identify 16 high-frequency traders using this definition, and conclude that while these traders did not trigger the flash crash, their responses exacerbated market volatility during the event. 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.

Several recent theoretical papers attempt to shed light on the potential impact of high-frequency trading in financial markets (Civitanic and Kirilenko, 2010; Gerig and Michayluk, 2010; Hoffmann, 2013; Jovanovic and Menkveld, 2010; Cohen and Szpruch, 2012; Biais, Foucault, and Moinas, 2012; Cartea and Penalva, 2012; Jarrow and Protter, 2012; Martinez and Rosu, 2013). Some of these papers have specific implications as to the relationships between high-frequency trading and liquidity or volatility, which we investigate empirically.

For example, Gerig and Michayluk (2010) assume that automated liquidity providers are more efficient than other market participants in extracting pricing-relevant information from multiple securities. By using information from one security to price another security, these high-frequency traders are able to offer better prices, lowering the transaction costs of investors in the market.

[29] Chordia, Roll, and Subrahmanyam (2011) look at recent trends in market activity. They argue that market quality, especially the efficiency of price formation, has improved in recent years as trading volume increased, possibly in part due to high-frequency trading.

Hoffmann (2013) introduces fast traders into the limit order book model of Foucault (1999). Their presence can (in some cases) lower transaction costs due to increased competition in liquidity supply. Cartea and Penalva (2012) construct a model in the spirit of Grossman and Miller (1988) except that they add high-frequency traders who 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.

In general, the theoretical models demonstrate that high-frequency traders can impact the market environment (and other investors) positively or negatively depending on the specific assumptions regarding their strategies and the assumed structure of the economy [see, for example, the predictions in Jovanovic and Menkveld (2010) and Biais, Foucault, and Moinas (2012)]. Since different types of proprietary algorithms may employ different strategies, a theoretical model that focuses on one strategy may shed light on the specific impact of such a strategy, but may not predict the overall effect that empirical studies find because the mixture of strategies in actual markets may overwhelm the effect of one strategy or the other. As such, while our results are more consistent with some models than others, we do not view them as necessarily suggesting that certain models are wrong. Rather, our results could point to the relative dominance of a subset of high-frequency traders who pursue certain strategies that improve market quality.

7. Conclusions

Our paper makes two significant contributions. First, we develop a measure of low-latency activity using publicly-available data that can be used as a proxy for high-frequency trading. Second, we study the impact that low-latency activity has on several dimensions of market quality both during normal market conditions and during a period of declining prices and heightened economic uncertainty. Our conclusion is that in the current market structure for equities, increased low-latency activity improves traditional yardsticks of market quality such as liquidity and short-term volatility. Of particular importance is our finding that at times of falling prices and anxiety in the market, the nature of the millisecond environment and the positive influence of low-latency activity on market quality remains. However, we cannot rule out the possibility of a sudden and severe market condition in which high-frequency traders contribute to 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. While our results suggest that market quality has improved, we believe it is as yet an unresolved question whether low-latency trading increases the episodic fragility of markets, and we hope that future research will shed light on this issue.

The millisecond environment we describe—with its clock-time periodicities, trading that responds to market events over millisecond horizons, and algorithms that "play" with one another—constitutes a fundamental change from the manner in which stock markets operated even a few years ago. Still, the economic issues associated with latency in financial markets are not new, and the private advantage of relative speed as well as concerns over the impact of fast traders on prices were noted well before the advent of our current millisecond environment.[30]

[30] Barnes (1911) describes stock brokers who, in the pre-telegraph era, established stations on high points across New Jersey and used semaphore and light flashes to transmit valuable information between New York and Philadelphia. He notes that some of the mysterious movements in the stock markets of Philadelphia and New York were popularly ascribed to these brokers.

The early advocates of electronic markets generally envisioned arrangements wherein all traders would enjoy equal access (e.g., Mendelson and Peake, 1979). 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.

The first impediment to equal access is the geographical dispersion of traders (Gode and Sunder, 2000). Our evidence on the speed of execution against improved quotes suggests that some players are responding within 2–3 ms, which is faster than it would take for information to travel from New York to Chicago and back (1,440 miles) even at the speed of light (about 8 ms). While co-location could be viewed as the ultimate equalizer of dispersed traders, it inevitably 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. Unless markets change from continuous to periodic, some traders will always have lower latency than others. It is of special significance, therefore, that our findings suggest that increased low-latency activity need not invariably work to the detriment of long-term investors in the post-Reg NMS market structure for U.S. equities.

References

Admati, A., Pfleiderer, P., 1988. A theory of intraday trading patterns: volume and price variability. Review of Financial Studies 1, 3–40.

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

Baum, C.F., Schaffer, M.E., Stillman, S., 2010. ivreg2: Stata Module for Extended Instrumental Variables/2SLS, GMM and AC/HAC, LIML and k-class Regression. ⟨http://ideas.repec.org/c/boc/bocode/s425401.html⟩.

Biais, B., Foucault, T., Moinas, S., 2012. Equilibrium high-frequency trading. Working paper. University of Toulouse.

Boehmer, E., Fong, K.Y.L., Wu, J.J., 2012. International evidence on algorithmic trading. Working paper. EDHEC Business School.

Boulatov, A., Dierker, M., 2007. Pricing prices. Working paper. University of Houston.

Brock, W., Kleidon, A., 1992. Periodic market closure and trading volume: a model of intraday bids and asks. Journal of Economic Dynamics and Control 16, 451–489.

Brogaard, J., 2012. Essays on High-Frequency Trading (Ph.D. dissertation). Northwestern University.

Brogaard, J., Hendershott, T.J., Riordan, R., 2012. High-frequency trading and price discovery. Working paper. University of California at Berkeley.

Carrion, A., 2013. Very fast money: high-frequency trading on NASDAQ. Journal of Financial Markets (in this issue).

Cartea, Á., Penalva, J., 2012. Where is the value in high-frequency trading? Quarterly Journal of Finance 2, 1250014.

Cespa, G., Foucault, T. Sale of price information by exchanges: does it promote price discovery? Management Science (forthcoming).

Chaboud, A., Hjalmarsson, E., Vega, C., Chiquoine, B., 2013. Rise of the machines: algorithmic trading in the foreign exchange markets. Working paper. Board of Governors of the Federal Reserve System.

Chordia, T., Roll, R., Subrahmanyam, A., 2011. Recent trends in trading activity and market quality. Journal of Financial Economics 101, 243–263.

Civitanic, J., Kirilenko, A.A., 2010. High-frequency traders and asset prices. Working paper. California Institute of Technology and MIT Sloan School.

Cohen, S.N., Szpruch, L., 2012. A limit order model for latency arbitrage. Mathematics and Financial Economics 6, 211–227.

Cragg, J.G., Donald, S.G., 1993. Testing identifiability and specification in instrumental variable models. Econometric Theory 9, 222–240.

Driscoll, J.C., Kraay, A.C., 1998. Consistent covariance matrix estimation with spatially dependent panel data. Review of Economics and Statistics, 549–560.

Easley, D., O'Hara, M., Yang, L., 2013. Differential access to price information in financial markets. Working paper. Cornell University.
Foucault, T., 1999. Order flow composition and trading costs in a dynamic limit order market. Journal of Financial Markets 2, 99–134.
Gerig, A., Michayluk, D., 2010. Automated liquidity provision and the demise of traditional market making. Working paper. University of Technology, Sydney.
Glosten, L.R., Milgrom, P.R., 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics 14, 71–100.
Gode, D.K., Sunder, S., 2000. On the impossibility of equitable continuously-clearing markets with geographically distributed traders. Working paper. Yale School of Management.
Grossman, S.J., Miller, M.H., 1988. Liquidity and market structure. Journal of Finance 43, 617–633.
Groth, S.S., 2009. Further evidence on "Technology and liquidity provision: the blurring of traditional definitions". Working paper. Goethe University.
Gsell, M., 2009. Algorithmic activity on Xetra. Journal of Trading 4, 74–86.
Gsell, M., Gomber, P., 2009. Algorithmic trading engines vs. human traders: do they behave different in securities markets. In: Proceedings of the 17th European Conference on Information Systems (ECIS), Verona, Italy.
Hagströmer, B., Norden, L.L., 2013. The diversity of high-frequency traders. Journal of Financial Markets (in this issue).
Hasbrouck, J., Saar, G., 2009. Technology and liquidity provision: the blurring of traditional definitions. Journal of Financial Markets 12, 143–172.
Hendershott, T.J., Jones, C.M., Menkveld, A.J., 2011. Does algorithmic trading improve liquidity? Journal of Finance 66, 1–33.
Hendershott, T.J., Moulton, P.C., 2011. Automation, speed and market quality. Journal of Financial Markets 14, 568–604.
Hendershott, T.J., Riordan, R., 2013. Algorithmic trading and the market for liquidity. Journal of Financial and Quantitative Analysis (forthcoming).
Hoffmann, P., 2013. Algorithmic trading in a dynamic limit order market. Working paper. Universitat Pompeu Fabra.
Jarrow, R.A., Protter, P., 2012. A dysfunctional role of high-frequency trading in electronic markets. International Journal of Theoretical and Applied Finance, 15.
Jovanovic, B., Menkveld, A.J., 2010. Middlemen in limit-order markets. Working paper. New York University.
Kirilenko, A.A., Kyle, A.S., Samadi, M., Tuzun, T., 2011. The flash crash: the impact of high frequency trading on an electronic market. Working paper. CFTC and University of Maryland.
Kosinski, R.J., 2012. A literature review on reaction time. ⟨http://biae.clemson.edu/bpc/bp/Lab/110/reaction.htm⟩.
Martinez, V.H., Rosu, I., 2013. High-frequency traders, news and volatility. Working paper. Baruch College and HEC Paris.
Mendelson, M., Peake, J.W., 1979. The ABCs of trading on a national market system. Financial Analysts Journal 35, 31–34 37–42.
Menkveld, A.J., 2013. High-frequency trading and the *new market* makers. Journal of Financial Markets (in this issue).
Moallemi, C.C., Sağlam, M., 2013. The cost of latency. Operations Research (in press).
Pagnotta, E.S., Philippon, T., 2013. Competing on speed. Working paper. New York University.
Prix, J., Loistl, O., Huetl, M., 2007. Algorithmic trading patterns in Xetra orders. European Journal of Finance 13, 717–739.
Riordan, R., Storkenmaier, A., 2012. Latency, liquidity and price discovery. Journal of Financial Markets 15, 416–437.
Stock, J.H., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Andrews, DWK, Stock, JH (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rotenberg. Cambridge University Press, Cambridge.
Thompson, S.B., 2011. Simple formulas for standard errors that cluster by both firm and time. Journal of Financial Economics 99, 1–10.
U.S. Commodities Futures Trading Commission and the U.S. Securities and Exchange Commission, 2010. Preliminary findings regarding the market events of May 6, 2010.
U.S. Securities and Exchange Commission, 2010. Concept release on equity market structure 34-61358.