JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS Vol. 54, No. 3, June 2019, pp. 993–1024
COPYRIGHT 2018, MICHAEL G. FOSTER SCHOOL OF BUSINESS, UNIVERSITY OF WASHINGTON, SEATTLE, WA 98195
doi:10.1017/S0022109018001096

Risk and Return in High-Frequency Trading

Matthew Baron, Jonathan Brogaard, Björn Hagströmer,
and Andrei Kirilenko*

Abstract

We study performance and competition among firms engaging in high-frequency trading (HFT). We construct measures of latency and find that differences in *relative* latency account for large differences in HFT firms' trading performance. HFT firms that improve their latency rank due to colocation upgrades see improved trading performance. The stronger performance associated with speed comes through both the short-lived information channel and the risk management channel, and speed is useful for various strategies, including market making and cross-market arbitrage. We find empirical support for many predictions regarding relative latency competition.

I. Introduction

Traditional models of market making argue that competition among market intermediaries should decrease their profits and lead to lower trading costs for other investors (Ho and Stoll (1983), Weston (2000)). Several models of high-frequency trading (HFT) adopt this standard view.[1] Other theories offer a contrasting perspective, that competition based on *relative* (i.e., rank-order) latency makes the HFT industry different and leads to a distinct competitive environment. For example, Foucault, Hombert, and Roşu (2016) and Foucault, Kozhan, and Tham (2017) show how competition based on relative latency can reduce market quality by increasing the adverse selection of non-HFT traders. Biais, Foucault, and Moinas (2015) and Budish, Cramton, and Shim (2015) find that

*Baron (corresponding author), baron@cornell.edu, Cornell University Johnson Graduate School of Management; Brogaard, brogaard@eccles.utah.edu, University of Utah Eccles School of Business; Hagströmer, bjorn.hagstromer@sbs.su.se, Stockholm University Business School; and Kirilenko, a.kirilenko@imperial.ac.uk, Imperial College Business School, Brevan Howard Centre for Financial Analysis. The authors thank Hank Bessembinder, Tarun Chordia, Thierry Foucault, Terry Hendershott, Charles Jones, Andrew Karolyi, Robert Korajczyk, Ananth Madhavan, Katya Malinova, Maureen O'Hara, Neil Pearson, Ryan Riordan, Gideon Saar, Ronnie Sadka, and Wei Xiong for their valuable feedback. The authors also thank Jennifer Conrad (the editor) and Ingrid Werner (the referee). The authors are grateful to Finansinspektionen for making data available for the article. Hagströmer is affiliated with the Swedish House of Finance and is grateful to the Jan Wallander and Tom Hedelius Foundation and the Tore Browaldh Foundation for research support.

[1] For example, Bongaerts and van Achter (2016), Jovanovic and Menkveld (2015), Aït-Sahalia and Saglam (2014), and Menkveld and Zoican (2017).

relative latency competition can lead to market concentration and an inefficient and costly technological arms race.[2] In their models, the fastest HFT firm responds first to profitable trading opportunities, capturing all the gains, while slower participants arrive marginally too late. As a result, small differences in trading speed are associated with large differences in trading revenues across firms, with trading concentrated among the fastest HFT firms.

Motivated by the view that competition based on relative latency differs from the competition among traditional market intermediaries, this article tests whether relative latency can explain cross-sectional differences in HFT performance. To our knowledge, we are the first to present direct evidence that small differences in trading speed are associated with large differences in trading revenues.

Although HFT firms benefit from the use of microwave transmission technologies (Shkilko and Sokolov (2016)) and colocation services (Brogaard, Hagströmer, Nordén, and Riordan (2015)), it is unclear to what extent speed matters for trading performance and through which channels. For example, Brogaard et al. (2015) find that not all HFT firms choose faster colocation technology when offered. Similarly, we find that only about half of the HFT firms react to market events at time scales near the latency frontier. This suggests that many HFT firms use computational power for other reasons, perhaps to better aggregate information from news feeds or order flow, and may not compete to be the fastest. Despite these alternative possibilities, we find that the HFT firms that are the fastest have better trading performance.

The theoretical literature has put forward a variety of channels through which HFT firms may translate speed into profitability. For example, HFT firms can use speed to enhance risk management, by avoiding adverse selection (Jovanovic and Menkveld (2015)) and improving inventory management (Aït-Sahalia and Saglam (2014)), or to trade on short-lived information (Foucault et al. (2016)). We find evidence that firms with lower relative latency are better along both of these dimensions. The fastest firms earn a higher realized spread when trading passively, consistent with better risk management. They also have the highest price predictability when trading with market orders, suggesting they are able to be the first to react to new information. Looking at cross-market arbitrage, we observe the fastest firms being more responsive to information on other exchanges. Speed is thus beneficial for a variety of strategies.

Our analysis uses proprietary transaction-level data with trading firm identifiers provided by the Swedish financial supervisory authority, Finansinspektionen. The data contain all trades of Swedish equities from Jan. 2010 to Dec. 2014 from all venues, including regulated exchanges, multilateral trading facilities, and dark pools. Given the high degree of fragmentation of volume in European equity trading, this cross-market coverage is an important feature to get the complete picture of trading. In addition, the 5-year length of our data is important in allowing us to

[2] In these models, the discontinuous difference in payoffs provides strong incentives to become marginally faster than other HFT firms through greater technological investment. Such competition on relative latency gives rise to a "positional externality" (Frank (2005)) because a firm that lowers its latency increases the *relative* latency of its competitors, which can in turn lead to an inefficient overinvestment in speed.

trace the “long-term” evolution of the HFT industry, at least relative to the rapid pace of innovation in the industry.

We focus on the 25 largest Swedish stocks by market capitalization. We classify HFT firms as those firms that self-describe as such through their membership in the Futures Industry Association’s European Principal Traders Association (FIA EPTA; a lobby organization for principal trading firms formed in June 2011) and any other firm that, according to its own Web site, undertakes low-latency proprietary trading. The 16 firms that we identify as HFT firms all have international trading operations, and none is headquartered in Sweden. Thus, it is unlikely that the findings reported in this article are specific to the Swedish context.[3,4]

We test the connection between HFT latency and trading performance. The main trading performance measure is REVENUES, captured daily for each HFT firm as the net of purchases and sales, marking end-of-day positions to market.[5] We also include risk-adjusted performance measures, including returns, factor model alphas, and Sharpe ratios. We find that HFT firms exhibit large, persistent cross-sectional differences in performance, with trading revenues disproportionally accumulating to a few firms. The results are robust to accounting for estimated exchange fees and liquidity rebates, which negligibly change the results.

Our main measure of latency is the difference in time stamps from a passive trade to a subsequent aggressive trade by the same firm, in the same stock and at the same trading venue. This measure, which we call DECISION_LATENCY, aims to capture the reaction time involved in a deliberate decision to trade in reaction to a market event (the HFT firm’s limit order being hit), which the HFT firm may view as informative. Specifically, for each HFT firm, we record the empirical latency distribution of all events where a passive trade is followed by an active trade by the same HFT firm in the same stock and at the same venue within 1 second. To capture the fastest possible reaction time for each HFT firm while also being robust to potential outliers, we use the 0.1% quantile of that distribution as the latency for each HFT firm. As an example of a strategy our measure may

[3] The data availability of the Swedish equity market has made it one of the most analyzed markets in the HFT literature. Hagströmer and Nordén (2013) show that HFT firms are highly active in this market, constituting approximately 30% of the trading volume and more than 80% of the order volume. Other empirical studies on this market include those by Breckenfelder (2013), Brogaard et al. (2015), Hagströmer, Nordén, and Zhang (2014), van Kervel and Menkveld (2019), and Menkveld and Zoican (2017).

[4] A previous version of this article analyzes HFT performance in the E-mini Standard & Poor’s (S&P) 500 futures contract over a 2-year period from 2010 to 2012. Whereas the E-mini is completely consolidated on one trading venue and has a relatively high relative tick size, Swedish equities trading is fragmented across multiple venues and features smaller relative tick sizes and lower trading volumes. Nevertheless, we generate similar findings (e.g., high industry concentration, difficulty of new entry, and the importance of latency), suggesting that the findings of this article are replicable, have external validity, and are robust to differences in market structure.

[5] Because our data set does not convey trading fees or other HFT operational costs, we are unable to directly calculate trading profits. However, in Table A5 of the Supplementary Material, we analyze the regulatory filings of five major HFT firms (Virtu, 2011–2015; Knight Capital Group, 2013–2015; GETCO, 2009–2012; Flow Traders, 2012–2015; and Jump Trading, 2010), which allows comparison of trading revenues and profits. We do not find evidence suggesting that higher trading revenues are associated with higher technological or operational costs and conclude that HFT revenue variation is a good proxy for variation in HFT profits. See the discussion in Section A6 of the Supplementary Material.

capture, Clark-Joseph (2013) shows that HFT firms use the execution of small "test" limit orders as a signal to trade on incoming order flow ahead of public order book feeds. Over our 5-year sample period, we show that the latency of the fastest HFT firms falls substantially.

We find that firms that are among the five fastest HFT firms, and in particular the fastest single HFT firm, earn substantially higher revenues than other HFT firms. Furthermore, we find that the fastest HFT firms capture more trading opportunities and have higher risk-adjusted revenues. We conclude, consistent with Biais et al. (2015) and Budish et al. (2015), that it is not being fast that allows an HFT firm to capture trading opportunities; it is being faster than competitors. Although the fastest-trading firms record higher trading volumes (trade quantity), we do not find any superior performance on a per-trade basis (trade quality). The differential finding suggests that the fastest HFT firms are no more accurate than other HFT firms at processing and analyzing information on a given individual trade, but their latency advantage allows them to seize more trading opportunities without taking on higher risk. The pattern that the fastest firms are responsible for the bulk of the trading volume is consistent with predictions by Roşu (2016).

Our results are robust to alternative approaches to measuring HFT latency. One alternative metric that we construct is QUEUING_LATENCY, which captures the race to be at the top of the order book. The measure is motivated by theoretical work by Yueshen (2014) and empirical findings by Yao and Ye (2018). Specifically, following price changes that lead to an empty price level in the limit-order book, we count how often a given HFT firm submits the first limit order and thus gets to the top of the queue. QUEUING_LATENCY does not rely on time stamps, making it robust to potential time-stamp noise. In addition, it is well suited to capture the latency of HFT firms that do not use market orders. Repeating the HFT performance analysis with the alternative latency metrics, our conclusions are unchanged.

To address possible endogeneity concerns, we present causal evidence from a quasi-experimental setting, studying two colocation upgrades on the NASDAQ OMX Stockholm exchange: the "Premium Colocation" upgrade first offered on Mar. 14, 2011, and the "10G Colocation" upgrade first offered on Sept. 17, 2012. These colocation upgrades lead some, but not all, HFT firms to get faster. We compare the change in trading performance for HFT firms that become relatively faster to those that become relatively slower. We show, as before, that increases in relative speed lead to better trading performance.

We then investigate through which channels relative latency benefits traders. As discussed previously, some theories view fast traders as using speed to trade on short-lived information, whether in reaction to news, order flow, or latency arbitrage. Other theories view speed as a way to avoid adverse selection and inventory costs. We proxy the short-lived information channel by the ability of a market order to predict price changes over the next 10 seconds, and we proxy the risk management channel by the ability of a passive order to capture a large realized spread. We find that relative latency is associated with better performance through both channels. As a specific strategy, we study cross-market arbitrage by examining HFT firms' equity trading following changes in the price of index futures. In the second after a change in the index futures price, the fastest HFT firms are more likely than other HFT firms to aggressively trade in individual equities in the direction of the futures price change. The fastest HFT firms are also less likely to supply liquidity to equities trades in the direction of the futures price change, which is consistent with avoiding adverse selection. We thus conclude that relative latency is important for performance both in short-lived information trading and in risk management.

We explore predictions regarding the effects of relative latency competition on market concentration. If the traditional view of market-making competition holds (Ho and Stoll (1983), Weston (2000)), we expect the alpha generated by HFT firms and the concentration of revenues to disappear as the industry matures. Alternatively, if HFT firms compete on relative latency, we do not expect increased competition to drive profit opportunities to zero. As argued by Budish et al. (2015), regardless of how fast the market as a whole becomes, there is always at least one firm with a relative speed advantage that can continue to adversely select other traders. Thus, rents should remain concentrated among the fastest HFT firms, even as overall market latency decreases.

Consistent with the predictions regarding competition on relative latency, we find that i) firm-level and industry-wide HFT performance is persistent; ii) HFT concentration of trading revenues and trading volumes is high and nondeclining over the 5-year sample, despite new HFT firm entry and a decline in overall HFT latency; and iii) new HFT entrants are typically slower, earn lower trading revenues, and are more likely to exit, which likely reinforces concentration in the HFT industry.

Finally, we find that the average cost of HFT activity to non-HFT traders ranges over time from 0.1 to 0.4 basis points (bps), which is on par with exchange taker fees and an order of magnitude lower than the effective spread for the stocks in our sample. We conclude that although HFT firms do compete on relative latency, the cost of the HFT industry incurred to other investors is low, at least on average or in normal market conditions. This conclusion is consistent with Brogaard et al. (2015) and Malinova, Park, and Riordan (2016), who provide empirical evidence that slow investors benefit from HFT firms through bid–ask spread reductions. Why might this be so? Our empirical results suggest that HFT firms utilize superior speed in a variety of ways that may partially offset each other. Furthermore, there are many HFT firms in our sample that are far from the cutting-edge in latency technology yet still have positive, albeit substantially lower, trading performance. The results suggest that HFT firms engage in a wider array of strategies than what the stylized theories of relative latency competition imply (Biais et al. (2015), Budish et al. (2015)). We discuss these issues further in Section VII.

The rest of the article is organized as follows: Section II presents the empirical framework, Section III characterizes HFT performance, Section IV analyzes the role of speed in HFT performance, Section V investigates through which channels latency impacts performance, Section VI explores the implications of latency competition for the HFT industry, Section VII discusses the results, and Section VIII concludes.

II. Empirical Framework

A. Data

Our primary data source is the Transaction Reporting System (TRS), a proprietary data set provided to us by Finansinspektionen, the Swedish financial supervisory authority. According to the Markets in Financial Instruments Directive (MiFID), financial institutions in the European Union that are under the supervision of one of the national financial supervisory authorities must report all their transactions with financial instruments to the TRS.

The TRS data have two features that make them highly suitable for the analysis of revenues in equity trading. First, the scope of the reporting obligation spans transactions at all trading venues, including regulated exchanges, multilateral trading facilities, and dark pools. This is important given the high degree of fragmentation of volume in European equity trading. Second, the TRS data contain identifiers (name, business identifier code, and address) for both the trading entity reporting the transaction and its counterparty. If the reporting entity undertakes the transaction as a broker for another financial institution, the identifiers for the client institution are reported too. The trading-firm identifiers are necessary to identify HFT firms and to analyze revenues in the cross section of firms. Finally, the TRS data contain standard transaction-level variables such as date, time, venue, price, currency, quantity, and a buy/sell indicator. See Section A1 of the Supplementary Material for information about the filtering procedures applied to the TRS data.[6]

We restrict the sample to the constituents of the leading Swedish equity index, the OMX S30, in order to focus on the most liquid stocks where HFT firms primarily operate (Hagströmer and Nordén (2013)). We exclude six stocks that are cross-listed in other currencies because revenue calculations for such stocks would require transaction data for foreign exchange markets.[7] There is one index constituent change during the sample period. We include Kinnevik Investment AB (KINVb) after its inclusion in the index on July 1, 2014, and we include Scania AB (SCVb) up until May 16, 2014, when it ceased trading. The final sample has 25 stocks covering the period Jan. 4, 2010–Dec. 30, 2014.

We match the TRS transactions to transaction-level data available from the Thomson Reuters Tick History (TRTH) database. The purpose of the matching is twofold. First, whereas the TRS data have second-by-second time stamps, the TRTH has time stamps at microsecond granularity. Through the matching, we can assign microsecond time stamps to the TRS data, which is important for our latency measurement. Second, the TRTH also contains order-book information synchronized to the transactions. This enables us to assess the status of the order book just before each TRS transaction, which is necessary to measure, for instance, the effective spread and to determine whether the trade was initiated by

[6] A limitation of the data set is that we cannot track activities in related securities, such as options and futures. To mitigate the effects this may have on inventory and revenue measurement, we exclude trades that are flagged in the data as derivative-related.

[7] The six stocks are ABB Ltd, Nokia Corporation, TeliaSonera AB, Nordéa Bank AB, AstraZeneca PLC, and LM Ericsson B.

the buyer or the seller, following Lee and Ready (1991).[8] The matching procedure is described in more detail in Section A1 of the Supplementary Material.

B. Institutional Detail

All the sample stocks have their primary listing at NASDAQ OMX Stockholm, which is open for continuous electronic limit-order-book trading from 9:00 AM to 5:25 PM on weekdays. For details about the trading mechanism at NASDAQ OMX Stockholm, see Hagströmer and Nordén (2013). Other important trading venues in our data are Chi-X, BATS, Turquoise, and Burgundy. In Feb. 2011, BATS and Chi-X merged at the corporate level, but they maintain separate trading venues throughout our sample period. Burgundy was acquired by Oslo Börs in 2012. All sample stocks are subject to mandatory central counterparty clearing.

Our sample stocks vary in market capitalization from 13,877 million SEK (MSEK) for SSABa to 475,595 MSEK for HMb, the equivalent of 1.78 to 60.91 billion USD (measured at closing prices on Dec. 31, 2014). In the U.S. equity market, stocks of that size are labeled as large-cap or mid-cap stocks. The more liquid stocks in our sample have turnover and bid–ask spreads similar to those of the US large-cap stocks studied by Brogaard, Hendershott, and Riordan (2014). For more descriptive statistics on our cross-section of stocks, see Section A2 of the Supplementary Material.

C. HFT Identification

Previous studies classify HFT firms according to observed trading behavior (following Kirilenko, Kyle, Samadi, and Tuzun (2017)) or using an exchange-defined classification (Brogaard et al. (2014)). We define HFT firms as those that self-describe as HFT firms by including firms that are members of the Futures Industry Association's European Principal Traders Association (FIA EPTA) or that, according to their own Web sites, primarily undertake low-latency proprietary trading. The advantage of this approach over a classification based on observed trading behavior is that we can verify that HFT firms have the characteristics usually associated with them: high trading volume, short investment horizons, and tight inventory management (U.S. Securities and Exchange Commission (SEC) (2010)).[9]

To include an HFT firm, we also require it to trade at least 10 MSEK a day (approximately 1.05 million USD at the exchange rate on Dec. 31, 2014) for at

[8] Concerns about the accuracy of the Lee–Ready algorithm (see Ellis, Michaely, and O'Hara (2000)) are mitigated by several features of this data set. First, trades inside the quotes are uncommon. This is due to the fact that the volume of hidden orders must exceed 50,000 euros, making such orders rare. There is a midpoint trading facility at NASDAQ OMX Stockholm, but its volume share is less than 0.1%. Second, misclassification due to fast trading is unlikely because for each trade recorded in the TRTH, there is also a quote update (usually with the same microsecond time stamp) reflecting how the trade influences the order book.

[9] As a robustness check, we alternatively use observed trading behavior to classify firms as HFT firms (i.e., if a firm has median daily trading volume of greater than 25 million SEK and median end-of-day inventory as a percentage of firm trading volume of less than 30%). The alternative specification addresses the possibility that some firms may not advertise themselves as HFT firms. Classification based on observed trading behavior produces nearly the exact same list of HFT firms as our main approach based on self-reporting.

least 50 trading days of the 1,255 trading days in the sample. We find 25 firms that self-describe as HFT firms, 16 of which satisfy the volume criteria and form our sample of HFT firms. The firm-day requirement of 10 MSEK is imposed to avoid outliers in trading performance that can appear due to small volumes. The nine firms that self-describe as HFT firm, but that do not satisfy the volume criteria together represent only 0.13% of the total HFT volume and 0.85% of the firm-day observations.[10]

D. HFT Performance Measures

We study three dimensions of performance: quantity measures, risk-adjusted measures, and quality measures. The quantity performance dimension measures the ability to capture trading opportunities, such as short-lived arbitrage events and the supply of liquidity to uninformed investors. The risk-adjusted performance dimension measures the ability to capture revenue while avoiding risky trades. The quality performance dimension measures the ability to capture revenues relative to trading volume.

We measure quantity performance using REVENUES and TRADING_VOLUME. REVENUES is defined as the cumulative cash received from selling shares, minus the cash paid from buying shares, plus the value of any outstanding end-of-day inventory positions marked to the market price at close.[11] We calculate REVENUES for each HFT firm, each sample stock, and each trading day. Depending on the application, we report REVENUES for different frequencies of time, for individual HFT firms as well as across all firms in the industry, and for individual stocks or all stocks; however, all versions of REVENUES are aggregates of the same panel of firm-stock-day observations. TRADING_VOLUME is the SEK volume traded, measured at the same frequency as REVENUES.

To capture risk-adjusted performance, we measure RETURNS, factor model alphas (1, 3, or 4 factors), and the SHARPE_RATIO. Through the use of risk-adjusted performance measures, we assess whether HFT firms with higher revenues are simply taking on more risk. The view that fast traders can achieve high risk-adjusted performance is supported by both theoretical models and real-world evidence. Aït-Sahalia and Saglam (2014) show that fast market makers are better at handling inventory risk, and Hoffmann (2014) shows that fast traders are able to avoid adverse-selection risk. In its initial public offering (IPO) prospectus, Virtu, an HFT firm in our sample, states: "we had only one losing trading day during …a total of 1,238 trading days."

RETURNS are calculated by dividing REVENUES for each firm by the implied capitalization of the firm. The implied capitalization is calculated for each HFT firm as the maximum position in SEK that a firm's portfolio takes over the 5-year sample. We find that HFT firms' inventories generally exhibit sharp,

[10] Due to confidentiality requirements, we cannot report the full list of names of the 25 HFT firms covered in the proprietary data set. However, in Appendix A3 of the Supplementary Material, we use public trading records to report the names of 19 HFT firms that trade at NASDAQ OMX Stockholm as members. The HFT firms not listed in Appendix A3 therefore trade only at other trading venues or as clients of other members at NASDAQ OMX Stockholm.

[11] See Appendix A4 in the Supplementary Material for a discussion and evaluation of alternative ways to account for inventory.

well-defined maximum and minimum total portfolio positions. We use the observed maximum position as an approximation of the maximum amount of capital that an HFT firm would need to execute its specific strategy in Swedish equity markets.[12] RETURNS can thus be viewed as the performance achieved relative to the capital allocated to the trading operation. RETURNS are calculated at daily frequencies but are reported in annualized values throughout the article.

Factor model alphas are computed for each HFT firm over the entire sample using the standard Fama–French (Fama and French (1993)) and Carhart (1997) momentum factors. The Fama–French and Carhart daily factors are constructed for Swedish equities according to the methodology from Fama and French (1993) and Ken French's Web site, using the full sample of Swedish stocks traded on NASDAQ OMX Stockholm. Methodological details concerning the construction of these factors and validation exercises are presented in Section A5 of the Supplementary Material.

The annualized SHARPE_RATIO for each HFT firm is calculated using daily observations as $\left(\mu_i - r_f/\sigma_i\right) \times \sqrt{252}$, where μ_i is the average daily return, r_f is the risk-free rate, and σ_i is the standard deviation of HFT firm i's returns.[13] We capture quality performance, the ability to enter trades with a high revenue margin, as REVENUES_PER_MSEK_TRADED. We calculate this measure daily as REVENUES divided by TRADING_VOLUME. None of the performance measures accounts for trading fees and liquidity rebates. We show in Section A7 of the Supplementary Material, however, that an adjustment for estimated exchange fees and liquidity rebates does not change the conclusions of the article.

E. HFT Latency

Generally, latency is the delay between a signal and a response, measured in units of time. Following Weller (2013), we define the signal as a passive execution for the HFT firm in question, and we define the response as a subsequent aggressive execution by the same firm. Examples of why HFT firms would attempt to trade aggressively immediately after a passive execution include "test" limit orders described by Clark-Joseph (2013) and "scratch" trades described by Kirilenko et al. (2017). An HFT firm cannot control the timing of the passive trade; it can only react to it. Our latency measure thus captures reactions to incoming order flow, not how fast an HFT firm can execute two successive trades.

Specifically, for each firm in each month, we record all cases where a passive trade is followed by an aggressive trade by the same firm, in the same stock, and at the same trading venue, within 1 second. By excluding cases where the two trades are recorded at different trading venues, we avoid potential problems related to the fact that time stamps may not be perfectly synchronized across venues. The time-stamp difference between the two trades in each case forms an empirical

[12] In Section A6 of the Supplementary Material, we show that HFT returns calculated this way are comparable in magnitude to those from the regulatory filings of five major HFT firms (Virtu, 2011–2015; Knight Capital Group, 2013–2015; GETCO, 2009–2012; Flow Traders, 2012–2015, and Jump Trading, 2010), where one can directly observe book capitalization or net liquid assets available to trade.

[13] Because the risk-free rate is effectively 0 during the sample period, the SHARPE_RATIO is calculated in practice as $\left(\mu(\text{REVENUES})_i/\sigma(\text{REVENUES})_i\right) \times \sqrt{252}$. Assumptions about equity capitalization are thus irrelevant for calculating the SHARPE_RATIO.

distribution of response times.[14] To capture the fastest possible reaction time while also being robust to potential outliers, we define DECISION_LATENCY as the 0.1% quantile of the aforementioned distribution.[15,16]

There are numerous signals that may trigger HFT firms to react swiftly, including news events, order-book gaps, and block orders. The inherent problem of signal-to-response latency measures is that HFT firms employ different strategies and assign different weights to different signals. We argue that it is likely that HFT firms respond to signals affecting their own portfolios, such as a passive execution. Our measure of DECISION_LATENCY, although not perfect, captures an important dimension of latency that varies across market participants. Although we conjecture that the passive trade is the information triggering the subsequent aggressive trade, this cannot be confirmed. Also, the measure is less informative for HFT firms that do not tend to follow passive executions with immediate aggressive executions.[17] However, these limitations should result in underestimating, not exacerbating, the role of speed in performance. Furthermore, in Section A10 of the Supplementary Material, we show that our results are robust to two alternative measures of latency.

[14] DECISION_LATENCY captures the following sequence of events: The starting point is when an HFT firm's resting limit order is executed by an incoming market order. The matching engine processes and time stamps the trade. A confirmation message is then sent to the HFT firm. The firm processes the confirmation information and makes a decision on how to react, which may be in the form of an aggressive order. The end of the latency measure is marked by the time stamp assigned when the message for the market order is processed by the matching engine.

[15] To ensure that DECISION_LATENCY is not picking up trades that happen close to each other by chance (or by time-stamp error that can also make time stamps randomly happen close to each other by chance), we simulate the probability of two successive trades, a passive trade followed by an aggressive trade, occurring by chance within a submillisecond interval. We find the probability to be small. Specifically, we simulate DECISION_LATENCY under the assumption that an HFT firm's trades within any venue or stock are uniformly distributed across a time period $[0,T]$; we then construct a simulated DECISION_LATENCY by examining the 0.1% quantile of the resulting latency observations of a passive trade followed by an aggressive trade. We make conservative assumptions: $T = 666{,}600$ trading seconds per month, and 37,431 aggressive trades and 59,162 passive trades per month, corresponding to the maximum observed aggressive and passive trades of any HFT firm in any stock–venue–month. Using simulation, we find the probability that DECISION_LATENCY is less than 50 microseconds to be less than 0.00001% for any firm–stock–venue–month observation. Given 15,169 firm–stock–venue–month observations in which HFT firms trade, the probability is less than $1 - (1 - 0.0000001)^{15169} = 0.2\%$ that even *one* of these 15.169 observations would be less than 50 microseconds by chance, even with these highly conservative assumptions. Thus, our empirical measurements of DECISION_LATENCY are almost certainly not due to chance or related to trading volume.

[16] The results are robust to using alternative quantile thresholds (0.5% and 1%) and MEAN_LATENCY, which is computed as the mean of this distribution conditional on being less than 1 millisecond. See Section A10 of the Supplementary Material.

[17] DECISION_LATENCY cannot be measured for HFT firms that trade exclusively using either aggressive or passive orders. Although 2.2% of the firm-months are subject to this limitation, those firm-months represent only 0.0007% of the trades. Another limitation of the DECISION_LATENCY definition is that fee differences may incentivize designated market makers (DMMs) to behave differently from other brokers. There are, however, no DMMs in our sample stocks.

III. Characterizing HFT Performance

We document the risk and return characteristics of individual HFT firms. In Table 1 we report the cross-sectional distribution of HFT performance, latency, and other trading characteristics. For each variable, we retrieve the time-series average for each HFT firm and then report the distributional statistics across firms.

The median HFT firm realizes on average REVENUES of SEK 6,990 per day, or SEK 56.5 REVENUES_PER_MSEK_TRADED. It has a daily TRADING_ VOLUME of MSEK 63.7, an annualized SHARPE_RATIO of 1.61, and a 4-factor (Fama–French plus Carhart momentum) annualized alpha of 9%. The RETURNS are also 9%, suggesting that exposure to well-documented risk factors is not particularly relevant for HFT firms.[18]

We find considerable performance variation in the cross-section of HFT firms. The cross-sectional distributions are skewed toward a few high performers. For example, firms in the 90th percentile generate REVENUES of SEK 61,354 per day, compared with SEK 6,990 for the median; a SHARPE_RATIO of 11.14,

TABLE 1

The Cross Section of HFT Performance

Table 1 reports descriptive statistics on high-frequency trading (HFT) performance and trading characteristics in the cross section of HFT firms. The following variables underlying the statistics are first aggregated for each day over all stocks and venues and then averaged across time for each HFT firm; the resulting cross-sectional distribution across HFT firms is then presented. REVENUES is the average daily trading revenue for each HFT firm, calculated as cash received from selling shares, minus the cash paid from buying shares, plus the value of any outstanding positions at the end-of-day marked-to-market price at close; TRADING_VOLUME is the average daily trading volume for each HFT firm, measured in MSEK; REVENUES_PER_MSEK_TRADED is daily REVENUES divided by daily TRADING_VOLUME for each HFT firm; RETURNS is daily REVENUES divided by the maximum intraday inventory position over the entire sample (used as a measure of capitalization) and reported in annualized terms; SHARPE_RATIO is the average monthly ratio for each HFT firm, reported in annualized terms, of the average daily return divided by the standard deviation of daily returns; the 1_FACTOR_ALPHA is the intercept estimated in a regression of daily HFT excess returns on the market return factor; the 3_FACTOR_ALPHA is the intercept estimated in a regression of daily HFT excess returns on the Fama–French factors; the 4_FACTOR_ALPHA is the intercept estimated in a regression of daily HFT excess returns on the Fama–French and Carhart momentum factors; END_OF_DAY_INVENTORY_RATIO is the absolute end-of-day SEK position (netted across stocks) divided by the TRADING_VOLUME; MAX_INTRADAY_INVENTORY_RATIO is the maximum absolute intraday SEK inventory position divided by the daily TRADING_VOLUME; INVESTMENT_HORIZON is the median holding time in seconds across all trades, calculated on a first-in, first-out basis; AGGRESSIVENESS_RATIO is the SEK volume traded using market orders divided by the TRADING_VOLUME; and DECISION_LATENCY is the 0.1% quantile of a distribution of latencies recorded in each firm-month where a passive trade is followed by an active trade at the same venue, in the same stock, within 1 second (measured in microseconds elapsed between the two trades of each event), averaged across months for each HFT firm. The sample consists of 25 Swedish stocks and 60 months of trading (Jan. 2010–Dec. 2014). $N = 16$ firms.

	Mean	Std. Dev.	P10	P25	P50	P75	P90
REVENUES (SEK)	18,181	29,519	−7,572	−487	6,990	31,968	61,354
REVENUES_PER_MSEK_TRADED	153.2	504.7	−257.9	−43.7	56.5	147.2	472.2
RETURNS	0.29	0.42	−0.09	0.01	0.09	0.51	0.89
SHARPE_RATIO	4.16	6.58	−1.47	0.33	1.61	7.02	11.14
1_FACTOR_ALPHA	0.29	0.43	−0.08	0.01	0.10	0.51	0.90
3_FACTOR_ALPHA	0.29	0.43	−0.07	0.01	0.09	0.51	0.94
4_FACTOR_ALPHA	0.29	0.43	−0.06	0.01	0.09	0.51	0.94
TRADING_VOLUME (MSEK)	272.0	378.1	4.2	7.4	63.7	507.7	909.2
END_OF_DAY_INVENTORY_RATIO	0.23	0.23	0.01	0.02	0.13	0.33	0.63
MAX_INTRADAY_INVENTORY_RATIO	0.28	0.25	0.03	0.07	0.18	0.41	0.70
INVESTMENT_HORIZON (seconds)	88.9	119.9	3.7	5.7	54.9	137.3	227.9
AGGRESSIVENESS_RATIO	0.51	0.26	0.16	0.28	0.56	0.69	0.88
DECISION_LATENCY (microseconds)	86,859	168,632	42	209	22,522	48,472	508,869

[18] Appendix A7 in the Supplementary Material analyzes HFT firm performance after accounting for potential maker-taker fees and liquidity rebates. Even after accounting for the most conservative possible fees and/or rebates, the trading performance for the entire distribution is shifted down slightly, but the results are qualitatively similar. For example, the performance results are still positively skewed, with the same HFT firms at the top strongly outperforming their competitors.

compared with 1.61 at the median; REVENUES_PER_MSEK_TRADED of SEK 472.2, compared with SEK 56.5 at the median; and a 4-factor annualized alpha of 94%, compared with 9% at the median.

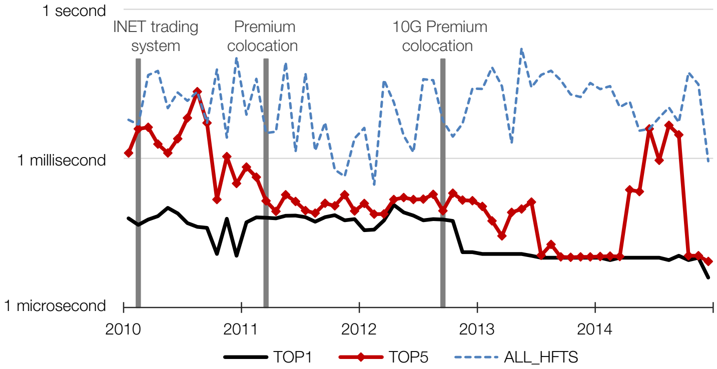
HFT firms are diverse in terms of other trading characteristics, too. We report the distributions of END_OF_DAY_INVENTORY_RATIO (the end-of-day inventory divided by TRADING_VOLUME), MAX_INTRADAY_INVENTORY_RATIO (the maximum intraday portfolio position divided by TRADING_VOLUME), INVESTMENT_HORIZON (the median holding time in seconds across all trades, calculated on a first-in, first-out basis); AGGRESSIVENESS_RATIO (the market-order volume in SEK divided by TRADING_VOLUME), and DECISION_LATENCY (in microseconds). Consistent with the characterization of HFT firms in the SEC's Concept Release on Equity Market Structure (2010) and with functional-based approaches for HFT classification (Kirilenko et al. (2017)), most, although not all, HFT firms tend to have low intraday and end-of-day inventories. HFT firms vary in their AGGRESSIVENESS_RATIO, with some trading almost exclusively actively or passively, whereas others use mixed order types. The average AGGRESSIVENESS_RATIO is 51%.

DECISION_LATENCY varies substantially across HFT firms, from 42 microseconds at the 10th percentile to 0.5 seconds at the 90th percentile. Notably, the 0.5-second latency for some HFT firms to process information and react with a market order is slow for automated traders but still fast relative to human reaction time. Given the important role of latency in our analysis, we also plot the time series of its distribution across HFT firms. In Figure 1, HFT firms are grouped by their relative rank of latency per month; the categories are TOP1, TOP5, and ALL_HFTS.

FIGURE 1

HFT Decision Latency over Time

Figure 1 plots DECISION_LATENCY (indicated on the vertical axis on a log scale) from Jan. 2010 to Dec. 2014. For each firm-month, DECISION_LATENCY is recorded as the 0.1% quantile of a distribution of latencies between a passive trade followed by an active trade at the same venue, in the same stock, within 1 second. TOP1 is the DECISION_LATENCY of the fastest high-frequency trading (HFT) firm in each month, TOP5 is the average DECISION_LATENCY of the 5 fastest HFT firms in each month, and ALL_HFTS is the average DECISION_LATENCY across all HFT firms for which latency is lower than 1 second in the given month. The vertical bars indicate microstructure events at NASDAQ OMX Stockholm that are expected to be associated with changes in latency. The sample consists of 25 Swedish stocks.



From 2010 to 2014, we find that latency decreases for the five fastest HFT firms, included in the TOP5 category. For example, the latency of the fastest HFT (TOP1) decreases from approximately 62 microseconds in 2010 to approximately 10 microseconds in 2014. (Note that this is not the same firm throughout the sample.) The relative reduction in latency is much greater for the TOP5 HFT firms, which start out in 2010 with latencies of over 1,280 microseconds and converge in latency to the TOP1 HFT by 2014.[19] In contrast, ALL_HFTS, which disproportionately picks up the slower HFT firms, remains relatively constant with an average latency of 25 milliseconds over the entire sample period. That HFT firms that are not among the five fastest do not achieve lower latencies over time is consistent with the findings of Brogaard et al. (2015). They show that not all HFT firms choose to be faster when given the opportunity to upgrade their colocation connection.

The magnitude of latency recorded for the fastest HFT firms in this article is in line with statements about the INET trading system used at NASDAQ OMX Stockholm. In marketing materials from 2012, NASDAQ states that its trading system delivers "sub-40 microsecond latency" (https://www.dropbox.com/s/tp2gjpcu9ge57r4/Nasdaq%20latency%20oct2012.pdf?dl=0). At that time, our fastest measured latency is approximately 60 microseconds.[20]

Figure 1 marks various technological upgrades: the introduction of INET in early 2010 (a high-capacity trading system, capable of handling over 1 million messages per second) and two colocation upgrades at NASDAQ OMX Stockholm in Mar. 2011 and Sept. 2012. Although it is difficult to assess the impact of the 2010 INET upgrade because it comes at the start of the sample, the colocation upgrade of 2012 is followed by a decline in latency for the TOP5 HFT firms. The fact that DECISION_LATENCY decreases following the technology upgrades provides suggestive evidence that our latency measure indeed captures reaction time. In Section IV, we use the 2011 and 2012 colocation upgrades to provide evidence on a causal relation between relative latency and trading performance.

IV. The Role of Speed in Performance

Having documented the performance and latency of HFT firms, we now test how the two are related. Although most theories posit that fast traders should have an informational advantage, other theories suggest that traders of different speed can specialize along other dimensions (Weller (2013), Roşu (2016)). According to these models, a relatively slow market intermediary could compensate by

[19] For the TOP5 group, there is a temporary increase in latency in mid-2014. This is driven by one of the preexisting top 5 firms withdrawing from the market and the former sixth-ranked HFT firm, which is substantially slower, joining the TOP5 group and raising the average. In Oct. 2014, a new, faster HFT firm joins the market.

[20] As additional points of reference, CME Globex advertised in Oct. 2015 *median* inbound latency of 52 microseconds, and the Swiss X-Stream INET exchange advertises *average* round-trip latencies of 33 microseconds for its ITCH Market Data interface. In 2015, the Bombay Stock Exchange claimed to operate the fastest platform in the world with a median response speed of 6 microseconds. It is important to note that these are median or average numbers, whereas we consider the 0.1% quantile.

providing deeper liquidity on the book or greater risk-bearing capacity, thus making similar profits as fast traders in equilibrium. Alternatively, some firms can simply be more skilled than others. For example, differences in technological capabilities can persist because technological expertise and trading strategies are closely guarded trade secrets, giving rise to barriers preventing the movement of human capital and technical knowledge across firms.

A. The Relation between Trading Performance and Latency

Motivated by the contrasting theories discussed in the Introduction, we test whether latency, and especially relative latency, is associated with increased performance.

We estimate the following regression model using ordinary least squares (OLS):

$$(1)\quad \text{PERFORMANCE}_{i,t} = \beta_1 \ln\left(\text{DECISION\_LATENCY}_{i,t}\right) + \beta_2 \text{TOP1}_{i,t} + \beta_3 \text{TOP5}_{i,t} + \gamma' \text{CONTROLS}_{i,t} + \text{month fixed effects} + \epsilon_{i,t},$$

where $\text{PERFORMANCE}_{i,t}$ is one of the HFT performance measures: REVENUES, RETURNS, SHARPE_RATIO, REVENUES_PER_MSEK_TRADED, or TRADING_VOLUME. All dependent variables are aggregated across stocks, venues, and days within the month to generate a firm-month panel on which equation (1) is estimated. Specifically, REVENUES and TRADING_VOLUME are averaged across trading days, and RETURNS and REVENUES_PER_MSEK_TRADED are calculated using the firm-month observations of REVENUES and TRADING_VOLUME. The factor model alphas are not included because they are nearly identical to RETURNS, as discussed in Section III.A.

Nominal latency enters the model as ln (DECISION_LATENCY). Because DECISION_LATENCY can vary widely across firms, from the microsecond to the second level (see Table 1), the relationship between trading speed and trading revenues is best captured by taking logs. Relative latency is represented by two indicator variables, $\text{TOP1}_{i,t}$ and $\text{TOP5}_{i,t}$, that flag whether a given firm is the fastest or among the five fastest HFT firms in a given month. Note that both indicators flag the fastest HFT firm each month. We design the indicator variables to capture the potentially nonlinear relationship between latency and performance: The fastest firms may perform substantially better than firms that are only slightly slower.

The control variables account for other characteristics that may affect HFT firms' performance, including measures of their risk-bearing capacity and trading strategies. These variables include the END_OF_DAY_INVENTORY_RATIO, MAX_INTRADAY_INVENTORY_RATIO, INVESTMENT_HORIZON, and AGGRESSIVENESS_RATIO, which are defined in Section III.A and calculated at a monthly frequency for each HFT firm. MAX_INTRADAY_INVENTORY_RATIO is used in the denominator to calculate RETURNS and is thus omitted when RETURNS is the dependent variable.

We normalize all continuous independent variables to be in units of standard deviations. Monthly fixed effects absorb time-varying market conditions, including market trading volume and volatility. Following Petersen (2009) and

Thompson (2011), standard errors are double clustered by firm and month to account for correlations both across firms and over time. Table 2 reports coefficient estimates for various specifications of the model described in equation (1).

Our first result is that being fast is associated with increased revenue. The first specification sets all slope coefficients except that of nominal latency (β_1) equal to zero and shows a negative and statistically significant relation between REVENUES and nominal latency.

The second result is that the effect of relative latency on REVENUES dominates that of nominal latency. This is seen in the second and third specifications, where the relative latency indicators ($\text{TOP1}_{i,t}$ and $\text{TOP5}_{i,t}$) are included along with the nominal latency variable in the second specification, and along with control variables in the third specification. The lack of statistical significance and reduced economic magnitude for ln(DECISION_LATENCY) suggest that relative speed matters more than nominal speed. Specifically, the estimates in column 3 of Table 2 show that being among the five fastest HFT firms (TOP5) predicts average daily trading revenues that are SEK 15,451 higher than those for firms outside the top five. Being the fastest (TOP1) provides on average daily trading revenues of SEK 24,639 in addition to the revenues from being among the five fastest. The t-test of the TOP1 coefficient shows that the difference in revenues between being TOP1 and TOP5 is statistically significant.[21] HFT firms that are not among the five fastest in a given month still have positive average daily revenues, amounting to SEK 10,894, as reflected by the intercept. The estimates imply that the fastest HFT firm has revenues approximately five times higher than those of the "slow" HFT firms.

Several of the control variables are related to trading revenues. For example, a standard-deviation increase in the MAX_INTRADAY_INVENTORY_RATIO is associated with decreased daily REVENUES of 21,008 SEK, suggesting that HFT firms that have tighter inventory management perform better. Similarly, HFT firms that are more aggressive earn somewhat higher trading revenues.

The results for latency effects on risk-adjusted performance measures are similar to those for REVENUES. The only difference between RETURNS and REVENUES is that the former is expressed relative to the firm market capitalization. The results thus indicate that the size of the HFT firm does not drive the relation between REVENUES and relative latency. Furthermore, we find that the SHARPE_RATIO is higher for HFT firms with lower relative latency. This demonstrates that the relation between REVENUES and relative latency is not driven by the risk of the trading strategies.

[21] Appendix A8 in the Supplementary Material repeats the analysis presented here but breaks down the TOP5 dummy variable into individual dummy variables for the fastest HFT firms: RANK1, RANK2, RANK3, RANK4, and RANK5. It shows that even among the top 5 HFT firms, the faster firm tends to perform better and that performance is monotonic in relative latency. Appendix A9 in the Supplementary Material repeats the analysis presented here with two different variations. First, we repeat the analysis including firm fixed effects. Second, we focus the analysis on HFT firms whose relative speed rank moves up and down. These additional analyses provide similar qualitative results, albeit with weaker statistical significance.

TABLE 2
Trading Performance and Latency

Table 2 analyzes the relationship between trading performance and latency. It reports coefficients estimated from equation (1) for 5 performance measures as dependent variables: REVENUES, RETURNS, SHARPE_RATIO, TRADING_VOLUME, and REVENUES_PER_MSEK_TRADED (all defined as in Table 1), which are calculated at a daily level for each high-frequency trading (HFT) firm aggregated across stocks and then averaged across trading days in each month to get firm-month observations (except for the SHARPE_RATIO, which is calculated as firm-month observations using the mean and standard deviation of daily observations of REVENUES aggregated across stocks). We estimate ordinary least squares (OLS) regressions with month fixed effects. The independent variables considered are as follows: ln(DECISION_LATENCY) is the natural logarithm of DECISION_LATENCY (defined as in Table 1). TOP1 and TOP5 are indicator variables for whether a given firm is ranked among the top 1 or top 1–5 firms by DECISION_LATENCY in a given month. END_OF_DAY_INVENTORY_RATIO, MAX_INTRADAY_INVENTORY_RATIO, INVESTMENT_HORIZON, and AGGRESSIVENESS_RATIO are defined as in Table 1. All continuous independent variables are in units of standard deviations. We omit MAX_INTRADAY_INVENTORY_RATIO as a control when estimating RETURNS as the dependent variable because MAX_INTRADAY_INVENTORY_RATIO is used in the denominator to calculate returns. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are double clustered by firm and month and are reported in parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (Jan. 2010–Dec. 2014).

	REVENUES			RETURNS			SHARPE_RATIO			TRADING_VOLUME ($\times 10^{-6}$)			REVENUES_PER_ MSEK_TRADED		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ln(DECISION_LATENCY$_{i,t}$)	14,020***	−1,063	9,925	−0.221***	−0.059	−0.00349	−4.38***	−1	2.03	−247***	−89.7	10.5	−19.4	−10.7	101**
	(4,311)	(6,358)	(10,481)	(0.0483)	(0.065)	(0.0852)	(0.632)	(1.2)	(1.46)	(43.7)	(59.1)	(74)	(57.5)	(69.1)	(40.4)
TOP1$_{i,t}$		29,849*	24,639**		0.238*	0.252*		3.77*	4.2*		326***	281***		6.99	57.6*
		(15,251)	(12,249)		(0.134)	(0.142)		(2.21)	(2.29)		(97.9)	(104)		(51.7)	(32.8)
TOP5$_{i,t}$		24,074**	15,451*		0.333**	0.303**		7.29**	5.61**		301**	201**		19.4	44.1
		(11,619)	(8,009)		(0.155)	(0.133)		(3.24)	(2.63)		(132)	(97.4)		(93.3)	(55.9)
END_OF_DAY_INVENTORY_RATIO$_{i,t}$			2,921			0.0839*			2***			−33.9**			326*
			(3,774)			(0.0494)			(0.74)			(15.9)			(168)
MAX_INTRADAY_INVENTORY_RATIO$_{i,t}$			−21,008**			*[omitted]*			−3.74***			−183***			−76.3
			(8,579)						(1.23)			(65.3)			(127)
INVESTMENT_HORIZON$_{i,t}$			−5,401			−0.134***			−2.25***			−76.4			−73.3
			(5,994)			(0.0404)			(0.726)			(50.3)			(63.8)
AGGRESSIVENESS_RATIO$_{i,t}$			5,481			−0.0212			−0.779			41.7			−55.5
			(3,865)			(0.0558)			(0.823)			(28.8)			(65.8)
Constant	20,278***	8,466**	10,894**	0.254***	0.104*	0.107**	5.1***	1.94	2.26*	313***	169***	198***	35.2	27	7.91
	(6,973)	(4,189)	(4,885)	(0.0579)	(0.0587)	(0.0513)	(1.26)	(1.23)	(1.23)	(75.9)	(57.8)	(56.3)	(57.3)	(80.2)	(10)
Month fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.123	0.168	0.263	0.198	0.233	0.269	0.207	0.254	0.361	0.294	0.362	0.454	0.080	0.080	0.148
N	737	737	737	737	737	737	737	737	737	737	737	737	737	737	737

To understand why relative latency is important, we next analyze whether it is primarily related to the trading revenues per trade (quality) or the number of trades (quantity). If HFT firms use latency advantages to better obtain and aggregate information in order to predict future price changes, we would expect the fastest HFT firms to have the highest revenues per trade. However, according to Table 2, we find only a weak statistical association between DECISION_LATENCY and REVENUES_PER_MSEK_TRADED, and when the control variables are included, the latency effects are not stable. Instead, we find a strong relationship between trading speed and TRADING_VOLUME. The results imply that the fastest HFT firms are no more accurate at processing new information per trade than other traders, but their latency advantage allows them to capture the most trading opportunities. This result supports the use of a measure of trade quantity, like REVENUES, rather than a measure of trade quality, like REVENUES_PER_MSEK_TRADED, when evaluating HFT performance.

B. Evidence from Two Colocation Upgrades

To address the potential endogeneity concern that another variable correlated with HFT latency might be driving trading performance, we put forward evidence from a quasi-experimental setting studying two colocation upgrades that cause some HFT firms to increase their relative speed. On Mar. 14, 2011, and Sept. 17, 2012, NASDAQ OMX Stockholm implemented optional upgrades to its colocation offerings (the "Premium Colocation" and "10G Colocation" upgrades, respectively). Members subscribing to the previously fastest colocation service were then offered to upgrade to an even faster connection. We study these events, which result in some HFT firms improving their latency rank, and find evidence in support of a causal relation between relative latency and trading performance.

The Sept. 17, 2012, colocation upgrade has been previously studied by Brogaard et al. (2015), and background and institutional detail on this event can be found in that article. In particular, they find that only about half of the affected members immediately subscribed to the new connection type.

Specifically, we measure DECISION_LATENCY before and after the event and compare the change in trading performance for HFT firms that become *relatively* faster through the colocation upgrade to HFT firms that become *relatively* slower. Given that DECISION_LATENCY consists of firm-month observations, we compare the latency of each firm the first full month before the colocation upgrade to the second full month after the event; for consistency, we measure the change in HFT performance for each firm over this same period. In measuring DECISION_LATENCY, we skip a month after the colocation upgrade because it seems that HFT firms take time to adopt and exploit this new technology; we observe that the distribution across firms of DECISION_LATENCY decreases and fully reaches a stable equilibrium by the second month. However, it is important to note that the horizon for assessing performance does not matter: the difference-in-difference results are robust to looking at the change in HFT performance in a 2-, 4-, 8-, or 12-week period before and after the upgrade.

We find two HFT firms for the Mar. 14, 2011, event and one HFT firm for the Sept. 17, 2012, event that improve their latency rank and refer to them as "Faster." We compare that group to three HFT firms for the Mar. 14, 2011, event and one

HFT firm for the Sept. 17, 2012, event that decline in latency rank, which we refer to as "Slower." All other HFT firms that have unchanged relative latency, including some that get nominally, but not relatively, faster, are excluded from this analysis.

Table 3 reports the trading performance measures for "Faster" and "Slower" HFT firms, before and after the colocation upgrade event. The results suggest that the group of HFT firms that improve their relative latency around the colocation upgrade ("Faster") also improve their trading performance. This result holds for all five measures of trading performance. The HFT firms in the "Slower" group also improve their REVENUES and REVENUES_PER_MSEK_TRADED, but less so than the "Faster" group. The "Slower" group also has a lower TRADING_VOLUME after the event, suggesting that those HFT firms capture fewer trading opportunities. As seen by the difference-in-difference estimate in the bottom line of Table 3, the "Faster" group improves relative to the "Slower" group in terms of all trading performance measures considered. For each performance measure, we test the null hypothesis that there is no difference in the before–after change between the "Faster" and "Slower" groups; the statistical significance of the difference-in-difference estimates is assessed with a t-test, where a p-value is computed under the null by taking the before–after changes in performance for each HFT firm as independent observations, pooling together the "Faster" and "Slower" groups. There are seven firms with changing relative latency, yielding 6 degrees of freedom.

The difference-in-difference estimates are large, positive, and often statistically significant due to the consistency and magnitude of the change, despite the

TABLE 3

Relative Latency and Trading Performance around Colocation Upgrades

Table 3 examines the relationship between trading performance and latency around two colocation upgrades on Mar. 14, 2011, and Sept. 17, 2012. Specifically, it analyzes two groups of high-frequency trading (HFT) firms that experience a change in their latency rank around each event: firms that improve their rank in terms of DECISION_LATENCY are in the "Faster" group, and firms that decline their rank are in the "Slower" group. The table then reports the change in average trading performance around the colocation events. The trading measures REVENUES, RETURNS, REVENUES_PER_MSEK_TRADED, and TRADING_VOLUME are calculated at a daily level for each HFT and then averaged across trading days and HFT firms for each period and group, and the SHARPE_RATIO is calculated using the mean and standard deviation of daily observations of REVENUES. The difference between the "before" and "after" periods are reported for each group and each variable. The bottom row of the table reports the difference-in-difference estimate between groups. We test the null hypothesis that there is no difference in the before–after change between the "Faster" and "Slower" groups; the statistical significance of the difference-in-difference estimates is assessed with a t-test, where a p-value is computed under the null by taking the before–after changes in performance for each HFT firm as independent observations, pooling together the "Faster" and "Slower" groups (there are $N=7$ firms with changing relative latency, yielding 6 degrees of freedom).

	REVENUES				RETURNS				SHARPE_RATIO			
HFT Latency Rank	Before	After	Diff.	Std. Error	Before	After	Diff.	Std. Error	Before	After	Diff.	Std. Error
Faster	9,537	52,770	43,233	(14,841)	0.022	0.158	0.136	(0.055)	1.47	1.68	0.20	(0.46)
Slower	31,557	32,811	1,255	(2,608)	0.777	0.748	–0.030	(0.067)	5.50	4.70	–0.81	(0.99)
Difference-in-difference			41,978***	(6,533)			0.165**	(0.045)			1.01*	(0.50)

	TRADING_VOLUME ($\times 10^{-6}$)				REVENUES_PER_MSEK_TRADED			
HFT Latency Rank	Before	After	Diff.	Std. Error	Before	After	Diff.	Std. Error
Faster	415.9	537.4	121.5	(101.8)	–21.3	87.7	109.0	(33.9)
Slower	448.6	398.1	–50.5	(43.1)	207.3	282.2	74.9	(65.2)
Difference-in-difference			171.9**	(50.1)			34.1	(40)

small sample size. Notably, the relative improvement is statistically significant and much stronger for the quantity measures and risk-adjusted measures than for the quality measure REVENUES_PER_MSEK_TRADED, which is not statistically significant. The findings are thus consistent with the evidence presented in the previous section.

Given the small sample of firms that get relatively faster or slower, the evidence presented in Table 3 should be seen as suggestive. Nevertheless, the results are consistent with the notion that improved relative latency leads to a boost in trading performance, in particular in terms of quantity performance measures.

C. Alternative Latency Measures

We acknowledge two potential concerns about the DECISION_LATENCY metric. First, measuring DECISION_LATENCY requires both limit and market orders, potentially discriminating against HFT firms that do not mix order types. Second, the microsecond time stamps reported by the TRTH are not assigned when the trading venue receives an order but when the information about the order arrives at the TRTH servers. Variation in the delay within venue can potentially introduce time-stamp noise in the DECISION_LATENCY metric, although we expect it to be mitigated by two factors: First, time-series variation in the delay is presumably stronger over longer time periods than within milliseconds, over which DECISION_LATENCY is measured; and second, variation across venues due to geographical distance does not influence DECISION_LATENCY, which only uses time stamps from within the same venue. Nevertheless, to address these concerns, we consider two alternative approaches to measuring HFT latency.

We construct two additional latency measures, QUEUING_LATENCY and MEAN_LATENCY, that address the problems just described. We reestimate the models in Table 2, and the results are qualitatively unchanged; see Section A10 of the Supplementary Material.

V. How Does Latency Impact Performance?

The theoretical literature identifies two channels through which traders benefit from being fast: short-lived information and risk management. The benefit of low latency through short-lived information is explored by Foucault et al. (2016). They show that fast traders trade aggressively on news, picking off stale quotes. Furthermore, Biais et al. (2015) and Foucault et al. (2017) show that fast traders can benefit from a superior ability to react to cross-market arbitrage opportunities. Chaboud, Chiquoine, Hjalmarsson, and Vega (2014) provide empirical evidence of fast traders pursuing cross-market arbitrage. The benefit of low latency in risk management is highlighted by Hoffmann (2014), who emphasizes that low latency allows liquidity providers to reduce their adverse-selection costs by revising stale quotes before they are picked off. Aït-Sahalia and Saglam (2014) add that fast traders can also benefit in terms of reduced inventory risk, which is supported empirically by Brogaard et al. (2015).

In this section, we investigate specifically how latency influences the two channels just discussed. We find that relative latency determines ability in both short-lived information trading and risk management.

A. Short-Lived Information and Risk Management in General

To capture the extent to which HFT trades predict future short-term price movements, we measure PRICE_IMPACT as the basis-point change in the bid–ask spread midpoint from just before a trade initiated by an HFT firm to 10 seconds after. To capture risk management in passive trading, we measure REALIZED_SPREAD as the basis-point difference between the transaction price and the bid–ask spread midpoint 10 seconds after a trade where an HFT firm is the liquidity provider. REALIZED_SPREAD captures the benefit of earning a wide bid–ask spread, as well as the ability to avoid supplying liquidity to trades with price impact. Each measure is calculated as the SEK-volume-weighted average across all trades of a given firm in each stock and each month. We interact both REALIZED_SPREAD and PRICE_IMPACT with a ± 1 buy–sell indicator variable, such that a higher coefficient corresponds to better trading performance. Because the analysis requires information on the bid–ask spread, we limit the measures to trades that can be matched to order-book data, both contemporary to the trade and 10 seconds later.

We reestimate equation (1) with PRICE_IMPACT and REALIZED_SPREAD as the dependent variables. As before, the $\text{TOP1}_{i,t}$ and $\text{TOP5}_{i,t}$ are indicators that capture relative latency, whereas ln(DECISION_LATENCY) captures nominal latency. The estimation is run with and without control variables.

Unlike in Table 2, the dependent variables considered here vary across stocks, implying a firm-stock-month data frequency. Accordingly, we add stock-level characteristics to the vector of control variables. We define QUOTED_SPREAD as the average bid–ask spread prevailing just before each trade, and we define the TICK_SIZE as the minimum price change, both expressed in basis points relative to the bid–ask spread midpoint. We define VOLATILITY as the average 10-second squared basis-point returns, and we define NONHFT_TRADING_VOLUME as the daily sum of SEK trading volume in each stock that does not involve HFT firms. Finally, we construct a FRAGMENTATION_INDEX as the inverse of the trading-volume Herfindahl index across the five largest trading venues (BATS, Burgundy, Chi-X, NASDAQ OMX Stockholm, and Turquoise), implying that the index is measured on a scale from 1 to 5. The stock-level characteristics are constant across HFT firms. Similarly, the control variables used in Table 2 are HFT firm characteristics, and we assign them the same value across stocks. All continuous independent variables are in units of standard deviations. Standard errors are double clustered by firm-stock and month. We present the results in Table 4.

As with HFT performance in general, we find for the specific channels that relative latency, not nominal latency, drives performance. The coefficients for the TOP5 relative latency dummy are statistically significant and economically meaningful for both PRICE_IMPACT and REALIZED_SPREAD. For example, being among the 5 fastest HFT firms increases PRICE_IMPACT by 0.645 bps, which can be related to the intercept of 3.960 bps. That means that the 5 fastest HFT firms have a price impact that is approximately 16% ($0.645/3.960 \approx 16\%$) higher than that of other HFT firms. In addition, the fastest HFT firm outperforms the price impact of other HFT firms by another 0.340 bps ($0.340/3.960 \approx 9\%$).

TABLE 4

Price Impact, Realized Spread, and Latency

Table 4 presents the relationship between DECISION_LATENCY and two dependent variables: active PRICE_IMPACT and passive REALIZED_SPREAD. PRICE_IMPACT is the basis-point change in the spread midpoint from just before to 10 seconds after a trade initiated by a high-frequency trading (HFT) firm. REALIZED_SPREAD is the basis-point difference between the transaction price and the bid–ask spread midpoint 10 seconds after a trade where an HFT firm was the liquidity provider. TOP1 and TOP5 are indicator variables for whether a given firm is ranked as the fastest or 1 of the 5 fastest firms in a given month. DECISION_LATENCY, END_OF_DAY_INVENTORY_RATIO, MAX_INTRADAY_INVENTORY_RATIO, INVESTMENT_HORIZON, and the AGGRESSIVENESS_RATIO are defined as in Table 1. We also control for the following stock-month-specific variables: TICK_SIZE (the average minimum price change relative the bid–ask spread midpoint), QUOTED_SPREAD (the average bid–ask spread prevailing just before each trade relative to the bid–ask spread midpoint), VOLATILITY (the average 10-second squared returns, calculated from bid–ask midpoints), FRAGMENTATION_ INDEX (the inverse of a Herfindahl index of trading volumes across the 5 largest trading venues), and NONHFT_TRADING_VOLUME (the SEK trading volume recorded by non-HFT traders). All continuous variables are in units of standard deviations. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are double clustered by firm-stock and month and are reported in parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (Jan. 2010–Dec. 2014); stock-firm-month observations for which an HFT firm does not trade actively or passively are excluded.

	PRICE_IMPACT		REALIZED_SPREAD	
	1	2	3	4
ln(DECISION_LATENCY$_{i,t}$)	−0.318 (0.225)	−0.494* (0.212)	−0.364*** (0.098)	−0.384*** (0.096)
TOP1$_{i,t}$	0.371* (0.201)	0.337* (0.193)	0.021 (0.131)	0.060 (0.126)
TOP5$_{i,t}$	0.730** (0.362)	0.645** (0.315)	0.448*** (0.136)	0.477*** (0.118)
END_OF_DAY_INVENTORY_RATIO$_{i,t}$		0.106 (0.158)		0.004 (0.066)
MAX_INTRADAY_INVENTORY_RATIO$_{i,t}$		0.256 (0.216)		0.087 (0.066)
INVESTMENT_HORIZON$_{i,t}$		−0.158 (0.182)		−0.047 (0.068)
AGGRESSIVENESS_RATIO$_{i,t}$		0.010 (0.121)		−0.054 (0.058)
NONHFT_TRADING_VOLUME$_{t,s}$		−0.100 (0.162)		0.145** (0.062)
VOLATILITY$_{t,s}$		−0.226 (0.161)		−0.462** (0.217)
FRAGMENTATION_INDEX$_{t,s}$		−0.162* (0.097)		−0.094* (0.054)
TICK_SIZE$_{t,s}$		−0.0469 (0.301)		−0.265* (0.136)
QUOTED_SPREAD$_{t,s}$		0.890** (0.379)		0.623*** (0.163)
Constant	3.910*** (0.182)	3.960*** (0.220)	−0.096 (0.084)	−0.108 (0.107)
Month × stock fixed effects	Yes	No	Yes	No
R^2	0.196	0.016	0.158	0.017
N	11,449	11,449	11,269	11,269

For REALIZED_SPREAD, the TOP5 coefficient is 0.477 bps, whereas the intercept is insignificantly different from 0. This indicates that being among the fastest HFT firms is important for being successful at the risk management required for passive trading. The results are robust to the inclusion of the control variables.

We conclude that relative latency is important both for improving active trading on short-lived information and for risk management in liquidity-provision strategies. This is consistent with theoretical models, such as that by Foucault et al. (2016) on active trading and those by Hoffmann (2014) and Aït-Sahalia and Saglam (2014) on passive trading.

B. Short-Lived Information and Risk Management in Cross-Market Arbitrage

In a more controlled environment, we reexamine both channels by focusing on cross-market trading between the futures market and equities. Specifically, we test if faster HFT firms are more likely than slower HFT firms to *actively* trade in equities in quick response to "news" in the futures market, where "news" is defined to be a price change in the OMXS30 futures above a certain size. We also ask whether faster HFT firms are less likely than slower HFT firms to be adversely selected in a passive trade in equity markets in response to "news" in the futures market. The investigation is in line with the theoretical setup of active fast trading by Biais et al. (2015) and Foucault et al. (2017).

We estimate the following probit regression, which in essence follows the setup of Hendershott and Riordan (2013) and Brogaard et al. (2015):

$$\Pr[\text{FAST\_HFT\_TRADES}] = \Phi[c + \beta\text{NEWS} + \gamma^{'}\text{controls} + \text{stock fixed effects}]. \tag{2}$$

The unit of observation is a trade. To capture which firms are trading quickly in response to "news" in the futures market, we consider equity market trades in the 1-second interval subsequent to a "news" event in the futures market. In that interval, the binary variable FAST_HFT_TRADES is 1 for trades executed by "Fast" HFT firms and 0 for trades by "Slow" HFT firms. We consider two alternative specifications of "Fast" HFT firms, corresponding to the TOP1 and TOP5 ranking variables used previously, based on DECISION_LATENCY. We estimate the probit model in equation (2) separately for the TOP1 and TOP5 specifications of FAST_HFT_TRADES. In both cases, a trade is defined as "Slow" when executed by an HFT firm that is not among the 5 fastest in that month. For the TOP1 specification, trades executed by HFT firms that are ranked from second to fifth are omitted. The β estimate can be interpreted as the increased probability of a "Fast" HFT firm trading in equities relative to a "Slow" HFT firm, in response to "news" in the futures market.

We define NEWS to be ±1 when the return on the OMXS30 futures during a 1-second window preceding the stock trade is large, defined as when the absolute return exceeds the top decile among nonzero absolute returns of that month, and 0 otherwise. NEWS takes the value +1 if the active party trades in the direction of the news, and –1 if in the opposite direction. This design implies that NEWS reflects any event that causes a large price change in the futures index. Note that all sample stocks are constituents of the index underlying the futures contract, making arbitrage activities between the two markets likely (Hasbrouck (2003)).

We control for the following variables, which may affect the probability of "Fast" HFT firms doing cross-market arbitrage. LAGGED_VOLATILITY is the average second-by-second squared return (multiplied by 1,000) over the previous 10 seconds; LAGGED_VOLUME is the SEK trading volume (divided by 100,000) over the previous 10 seconds; QUOTED_SPREAD is defined as before; and DEPTH_AT_BBO is the average number of shares available at the best bid quote and the best offer quote (divided by 100,000), multiplied by the bid–ask spread midpoint.

Estimates for active trading (all trades initiated by an HFT market order) are reported in Panel A of Table 5. Similarly, the estimates for passive trading (all trades where an HFT firm supplies liquidity) are presented in Panel B. For computational tractability, we estimate the model separately for each month in 2010–2014. Similar to the Fama–MacBeth (1973) procedure, we then

TABLE 5
Cross-Market Arbitrage and Latency

Table 5 reports probit regression estimates corresponding to the probability of initiating a trade (active trading; Panel A) or supplying liquidity in a trade (passive trading; Panel B) in the equity markets in response to a change in the futures index price. The dependent variable is 1 when a "Fast" high-frequency trading (HFT) firm performs the trade, and it is 0 if a "Slow" HFT firm does it. In each month, "Slow" HFT firms are those that are not among the top 5 HFT firms in terms of trading speed. "Fast" HFT firms are either the fastest (TOP1) or among the 5 fastest (TOP5) HFT firms in terms of trading speed in the month of the trade. Trades included in the analysis are those performed by HFT firms in the sample stocks. A news event is defined as when the absolute return on the OMXS30 futures during a 1-second window preceding the stock trade is "large" (in the top decile for each month among nonzero absolute returns). The variable NEWS is +1 if the active party trades in the direction of the news, −1 if the active party trades in the opposite direction of the news, and 0 if there is no news event in the 1-second window before the trade. We include the following control variables: LAGGED_VOLATILITY, the average second-by-second squared return (multiplied by 1,000) over the previous 10 seconds; LAGGED_VOLUME, the SEK trading volume (divided by 100,000) over the previous 10 seconds; QUOTED_SPREAD, the difference between the best bid and offer quotes (multiplied by 10,000), divided by the midpoint quote; and DEPTH_AT_BBO, the average number of shares available at the best bid quote and the best offer quote (divided by 100,000), multiplied by the midpoint quote. Marginal effects are also reported. Regressions are estimated month by month from Jan. 2010 to Dec. 2014; reported coefficients and marginal effects are means across the 60 months, and standard errors are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	"Fast" =TOP1		"Fast" =TOP5	
	Probit (1 = "Fast" HFT)	Marginal Effects	Probit (1 = "Fast" HFT)	Marginal Effects
Panel A. Active Trading				
Constant	1.055***		2.143***	
	(0.31)		(0.17)	
NEWS	0.139***	0.006	0.199***	0.008
	(0.04)		(0.03)	
LAGGED_VOLATILITY	−0.094	0.000	−0.007***	−0.001
	(0.08)		(0.00)	
LAGGED_VOLUME	−0.005***	0.000	−0.004***	0.000
	(0.00)		(0.00)	
QUOTED_SPREAD	−0.046***	−0.004	−0.035***	−0.001
	(0.01)		(0.00)	
DEPTH_AT_BBO	0.013	0.006	0.049***	0.001
	(0.03)		(0.02)	
Stock fixed effects	Yes		Yes	
Avg. N	109,684		277,044	
Avg. pseudo-R^2	0.209		0.169	
Panel B. Passive Trading				
Constant	0.551*		1.646***	
	(0.30)		(0.15)	
NEWS	0.001	0.004	−0.097***	−0.015
	(0.03)		(0.02)	
LAGGED_VOLATILITY	−0.116	0.001	0.008***	0.001
	(0.12)		(0.00)	
LAGGED_VOLUME	0.001	0.000	0.000	0.000
	(0.00)		(0.00)	
QUOTED_SPREAD	−0.024	−0.002	−0.001	0.000
	(0.02)		(0.00)	
DEPTH_AT_BBO	−0.120**	−0.009	−0.015	−0.003
	(0.05)		(0.02)	
Stock fixed effects	Yes		Yes	
Avg. N	95,268		258,409	
Avg. pseudo-R^2	0.204		0.163	

average the coefficients across months to produce the estimates reported in Table 5. We also report marginal effects that show the increased probability of "Fast" HFT firms to engage in a trade if the explanatory variable increases by one standard deviation, conditional on all other explanatory variables being at their unconditional means.

We find that the NEWS coefficients are positive and statistically significant for active trading (Panel A of Table 5). Based on the marginal effects, the fastest HFT firm (TOP1) is 0.6% more likely to actively trade in equities subsequent to "news" arrival in the futures market, relative to a "Slow" HFT firm. The results for the 5 fastest HFT firms are similar. Overall, we conclude that faster HFT firms are more likely to quickly submit market orders in response to changes in the futures index. This is consistent with fast active traders being better positioned to pursue cross-market arbitrage, as modeled by Biais et al. (2015) and Foucault et al. (2017).

For passive trading, the NEWS coefficients are negative (Panel B of Table 5), indicating that fast HFT firms are less likely to get caught in passive equities trades that incur adverse-selection costs to the liquidity provider. The NEWS coefficient is not significantly different from 0 for the TOP1 specification, but for the TOP5 group, it is negative and significant at the 1% level. This is in line with Foucault et al. (2017), who find that the probability of toxic arbitrage is related to the latency of arbitrageurs relative the latency of liquidity providers.

VI. Implications of Latency Competition for the HFT Industry

The evidence presented previously indicates that HFT firms compete on relative speed. According to theoretical models, such competition leads trading revenues to be concentrated among the fastest participants. Budish et al. (2015) show that profits are not competed away under relative latency competition because regardless of how fast the market as a whole becomes, there is always at least one firm with a relative speed advantage that can continue to adversely select other traders. The prediction is that both firm-level and industry-wide performance is persistent. Furthermore, relative latency competition predicts barriers to entry for potential competitors, due to the difficulty for new entrants to immediately achieve a high ranking. In this section, we explore some of the predicted consequences of competition on relative latency.[22]

A. Market Concentration

We assess market concentration using a Herfindahl index of REVENUES:

$$(3)\quad \text{REVENUES\_CONCENTRATION}_t = \sum_{i=1}^{N_t}\left[\frac{\text{REVENUES}_{i,t}}{\sum_{i=1}^{N_t}\text{REVENUES}_{i,t}}\right],$$

[22] Other articles that study competition in the HFT industry include those by Boehmer, Li, and Saar (2018), who study competition between HFT firms within 3 distinct strategies and show that increased competition is associated with lower volatility and the migration of trading volume to newer venues; and Brogaard and Garriott (2019), who analyze the entry and exit of HFT firms and show that increased HFT competition increases market liquidity.

where N_t is the number of HFT firms that earn nonnegative trading revenues in month t, and REVENUES$_{i,t}$ is the revenues of firm i in each 6-month period t. We construct a similar index for TRADING_VOLUME_CONCENTRATION. We plot the time series of each concentration index in Graph A of Figure 2. The level of concentration, ranging from 0.186 to 0.304 for trading volume and from 0.275 to 0.354 for revenues, is comparable to what Van Ness, Van Ness, and Warr (2005) find for market makers at NASDAQ (a range of 0.037–0.439). Notably, the level of concentration is virtually constant over our 5-year sample.[23] The nondeclining concentration is consistent with competition on relative speed.

B. Persistence in Performance

We calculate industry-wide performance measures by aggregating across firm-level measures. Graph B of Figure 2 plots biannual measures of TOTAL_DAILY_REVENUES (REVENUES summed across all HFT firms). Graph C presents time series for HFT_REVENUES_PER_FIRM (the TOTAL_DAILY_REVENUES averaged across HFT firms), SEK_TRADING_VOLUME_PER_FIRM (the TRADING_VOLUME averaged across HFT firms and reported in MSEK), and REVENUES_PER_MSEK_TRADED (the ratio of HFT_REVENUES_PER_FIRM and SEK_TRADING_VOLUME_PER_FIRM). We find that TOTAL_DAILY_REVENUES increases slightly over our sample period.[24] Interestingly, the SEK_TRADING_VOLUME_PER_FIRM trends up while REVENUES_PER_MSEK_TRADED trends down, but the ratio of the two, HFT_REVENUES_PER_FIRM, is stable. One possible interpretation is that although HFT firms are competing more by increasing trading volume and pursuing ever-lower latencies, they are chasing the same number of profit opportunities, so the resulting HFT revenues per firm is the same.

To analyze persistence for individual HFT firms, we regress monthly firm-level performance measures (REVENUES, RETURNS, SHARPE_RATIO, and REVENUES_PER_MSEK_TRADED) on their lagged values (see Section A11 of the Supplementary Material for details about the regressions and for tabulated results). The results are reported in Table A13 of the Supplementary Material. We find that HFT firms have statistically significant monthly persistence coefficients: 0.631 for REVENUES, 0.763 for the SHARPE_RATIO, and 0.446 for RETURNS (where 1 indicates maximum persistence, and 0 indicates no persistence). Firms that have done well in the past typically continue to outperform their competitors. Consistent with our earlier argument that REVENUES_PER_MSEK_TRADED may be a less relevant performance metric for HFT firms, we find no significant persistence in this measure.[25] The evidence of persistence in performance for individual HFT firms as well as for the industry as a whole is consistent with the predictions of relative latency competition.

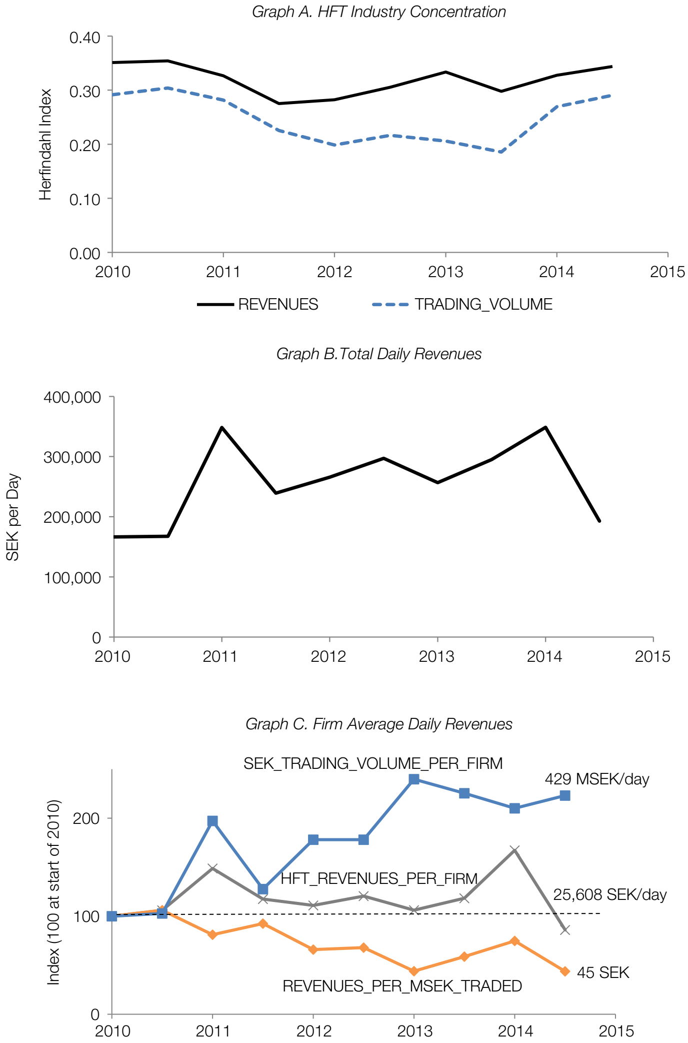
[23] In Appendix A11 of the Supplementary Material, we test for statistically significant time trends in the market concentration indexes. We find no significant time trend for revenue concentration, but we find a slight downward trend for volume concentration.

[24] Again, we test the significance of time trends in Section A11 of the Supplementary Material. All trends reported in this paragraph are statistically significant. The exception is HFT_REVENUES_PER_FIRM, which has no significant trend.

[25] Detailed results, including daily frequency performance measures and rank-order regressions, are presented in Section A11 of the Supplementary Material.

FIGURE 2
Time Series of the HFT Industry

Figure 2 shows 5-year time series for the high-frequency trading (HFT) industry. Graph A plots two indexes of industry concentration, calculated on daily REVENUES or with daily TRADING_VOLUME (see equation (3)). Graph B plots TOTAL_DAILY_REVENUES, which is the average daily sum of REVENUES (defined as in Table 1) across all firms. Graph C plots FIRM_AVERAGE_DAILY_REVENUES, which is the average REVENUES across all HFT firms active on each given day. This variable is equivalent to SEK_TRADING_VOLUME_PER_FIRM times REVENUES_PER_MSEK_TRADED, both of which are also plotted. The sample consists of 25 Swedish stocks and 60 months of trading (Jan. 2010–Dec. 2014). The time series are reported on a biannual frequency.



Graph A. HFT Industry Concentration

Graph B.Total Daily Revenues

Graph C. Firm Average Daily Revenues

C. Barriers to Entry

Potential barriers to entry include frictions in labor markets that may prevent the movement of human capital and technical knowledge to new HFT firms. For instance, technological expertise and trading strategies are closely guarded trade secrets, and employees often agree to noncompete and nondisclosure agreements.

If new entrants cannot simply pay to acquire human capital or the latest trading technology but can only acquire them from experience, then they may earn less and be less likely to survive in the market.

To evaluate the importance of experience, we regress performance measures on indicator variables for the length of time an HFT firm i has been active in a given stock s, designating less than 1 month, 2 months, or 3 months as indicators of new entry (denoted M1, M2, and M3). We estimate the following regression model using OLS:

$$(4) \qquad \text{PERFORMANCE}_{i,t,s} = \beta_1 \text{M1}_{i,t,s} + \beta_2 \text{M2}_{i,t,s} + \beta_3 \text{M3}_{i,t,s} + (\text{day} \times \text{stock})\text{ fixed effects} + \epsilon_{i,t,s},$$

where $\text{PERFORMANCE}_{i,t,s}$ can be REVENUES, REVENUES_PER_MSEK_TRADED, or RETURNS and is defined on a firm-stock-day frequency over the period 2010–2014. We exclude the observations in the first 3 months of the sample because new entry during this period cannot be established. The dummy variable $\text{M1}_{i,t,s}$ is equal to 1 if firm i began trading stock s in the last 30 calendar days, and 0 otherwise. Similarly, the $\text{M2}_{i,t,s}$ dummy corresponds to beginning trading in stock s in the last 31–60 days, and the $\text{M3}_{i,t,s}$ dummy corresponds to the previous 61–90 days.[26] The model estimates are reported in Table 6. Standard errors are double clustered by firm-stock and month and are reported within parentheses.

New entrants have significantly lower REVENUES and RETURNS than incumbent HFT firms. For REVENUES, the coefficient estimates are negative and significant for all three new-entry dummies. However, for REVENUES_PER_MSEK_TRADED, all three new-entry indicators are statistically insignificant. The fact that the new firms have lower total trading revenues but no statistically significant difference in REVENUES_PER_MSEK_TRADED is consistent with HFT firms competing on quantity, not quality, and new firms being less able to compete on capturing quantity.

Additionally, we examine whether new entrants are more likely to exit the market by estimating the model described in equation (4) with the dummy $\text{EXIT}_{i,t,s}$ as the dependent variable, which is equal to 1 on the last day t when firm i trades stock s, and 0 otherwise. Furthermore, we investigate whether new entrants have higher latency than incumbents by analyzing the same model with DECISION_LATENCY as the dependent variable.

The results in Table 6 show that new entrants have a higher probability of exiting compared with more established HFT firms. The daily probability of exit is significantly higher in a firm's first 2 months of trading. To see the economic significance of the result, note that the coefficient 1.455 implies that the probability of exit is 1.455% higher on *each day* of the first month. The latency analysis shows that new entrants struggle not only in performance but also in trading speed.

[26] In accounting for entry and exit, we ignore gaps and count the overall first and last trading days of a firm in a stock as entry and exit dates. HFT firm mergers (of which there are 2 in our sample) are also not counted as entry/exit events.

TABLE 6
HFT Entry and Exit Analysis

Table 6 analyzes the determinants of high-frequency trading (HFT) firm entry and exit into stocks. The performance measures REVENUES, REVENUES_PER_MSEK_TRADED, and RETURNS are defined as in Table 1 for each HFT firm, stock, and trading day. The table reports coefficient estimates from equation (4), estimated on a panel of firm-stock-day observations. The 1-month dummy M1 takes the value of 1 if firm *i* began trading stock *s* in the last 30 days, and 0 otherwise; 2- (M2) and 3-month (M3) dummy variables are defined similarly. We exclude the observations in Jan. 2010 because this is when we first observe any firm. In the fourth column, a linear probability regression is estimated using equation (4) but with the dummy $\text{EXIT}_{i,t,s}$ as the dependent variable, which takes the value of 1 on day *t* for firm *i* if that is the last day firm *i* trades stock *s*. This regression excludes observations in Dec. 2014, the last month of the analysis, because we cannot determine exits. In the fifth column, an ordinary least squares (OLS) is estimated on a firm-stock-month panel using equation (4) but with DECISION_LATENCY as the dependent variable. (Because DECISION_LATENCY is a firm-month variable, it is assigned to be constant across stocks). *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are double clustered by firm-stock and month and are reported in parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (Jan. 2010–Dec. 2014).

	REVENUES (thousands of SEK)	REVENUES_ PER_MSEK_ TRADED	RETURNS	Daily Probability of Exit (×10^3)	DECISION_ LATENCY (millisecond, monthly obs.)
$\text{M1}_{i,t,s}$	−1.90**	−97.46	−0.032**	1.455***	44.36*
	(0.93)	(209.7)	(0.014)	(0.420)	(26.73)
$\text{M2}_{i,t,s}$	−3.05**	−87.11	−0.033***	1.486***	134.6***
	(1.434)	(230.3)	(0.011)	(0.425)	(33.98)
$\text{M3}_{i,t,s}$	−0.78	104.7	−0.018*	−0.194	22.8**
	(1.35)	(196.6)	(0.010)	(0.477)	(11.19)
Constant	1.43***	76.64***	0.017***	.530***	14.87***
	(0.19)	(4.12)	(0.002)	(0.049)	(0.89)
Fixed effects	Day × stock	Day × stock	Day × stock	Day × stock	Month × stock
R^2	0.101	0.129	0.147	0.154	0.432
N	241,053	241,053	241,053	241,053	11,014

We find statistically significant coefficients on the 1-month, 2-month, and 3-month dummies of 44.36, 134.6, and 22.8 milliseconds, respectively.[27]

A high and steady industry concentration combined with strong firm-level persistence of REVENUES and RETURNS and the difficulty of new entry suggests that top-performing incumbent HFT firms maintain their position in the market. Together, the measures of industry structure we examine point to high and nondeclining concentration, consistent with relative latency driving performance.

VII. Discussion

The empirical evidence in Sections IV and V of this article shows that HFT firms compete on relative latency. The prior literature has divergent viewpoints regarding the implications of relative latency competition on market quality. On the one hand, recent theory has put forward concerns about relative latency competition, including overinvestment in technology (Biais et al. (2015), Budish et al. (2015)) and adverse-selection costs on slower market participants (Foucault et al. (2016), Foucault et al. (2017)). On the other hand, there is empirical evidence

[27] Our definition of DECISION_LATENCY does not allow for variation across stocks. For the purpose of this regression, we construct a firm-stock-day panel, where all observations across stocks for a given firm-day are assigned the same DECISION_LATENCY value.

that fast trading reduces the indirect costs of trading for slow traders. For example, Brogaard et al. (2015) show that the colocation upgrade at NASDAQ OMX Stockholm reduced the effective spread paid by trading firms *without* colocation by 2.8%.[28] Malinova et al. (2016) find that high-frequency market making in the Canadian market reduces the bid–ask spread for all traders, and in particular for retail investors.

To help assess these competing viewpoints, we define the COST_OF_HFT_INTERMEDIATION_FOR_NONHFTS as HFT firm TOTAL_DAILY_REVENUES divided by the trading volume (measured in SEK) associated with non-HFT traders. The measure captures the amount of revenue paid from non-HFT traders to HFT firms per SEK traded. We calculate biannual averages and plot the time series in Figure 3.

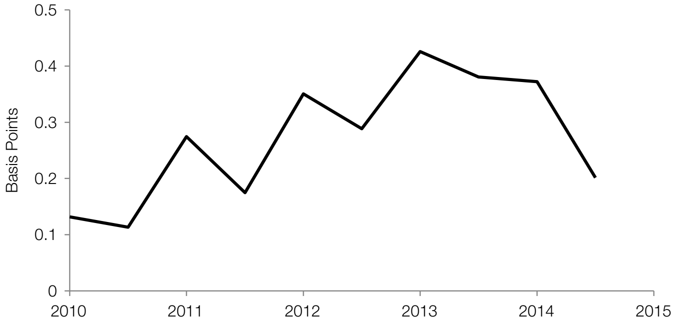
We find that the COST_OF_HFT_INTERMEDIATION_FOR_NONHFTS ranges from 0.113 to 0.426 bps. The cost is thus on par with exchange-taker fees that range from 0.15 to 0.50 bps (see Section A7 of the Supplementary Material). It is an order of magnitude smaller than the effective spread, which varies between 2 and 6 bps for our sample stocks (see Section A1 of the Supplementary Material). Thus, our results show that despite the potential downsides associated with relative latency competition, the cost of the HFT industry incurred to other investors is low, at least on average or in normal market conditions.

We see two reasons that competition on speed does not lead to high costs for the end investors. First, our empirical results suggest that HFT firms utilize superior speed in a variety of ways that may partially offset each other. For example, although we find that the fastest HFT firms are indeed better able to capture latency arbitrage opportunities, predict short-term price movements, and engage in cross-market arbitrage with aggressive trading, we also find that the fastest HFT firms are better positioned to manage adverse-selection risk in passive

FIGURE 3

Cost of HFT Intermediation to Non-HFT Traders

Figure 3 shows a 5-year time series of COST_OF_HFT_INTERMEDIATION_FOR_NONHFTS, calculated as high-frequency trading (HFT) firm TOTAL_DAILY_REVENUES (defined as in Figure 2) divided by the total NONHFT_TRADING_VOLUME (in SEK). The sample consists of 25 Swedish stocks and 60 months of trading (Jan. 2010–Dec. 2014).



[28] The evidence given by Brogaard et al. ((2015), p. 3412) shows that this was because it was the liquidity providers who chose to upgrade.

trading (both in the general context and in cross-market arbitrage). The evidence is consistent with that of Hoffmann (2014), who predicts that successful risk management in passive trading allow fast traders to quote tighter spreads. In addition, Brogaard et al. (2015) present evidence in support of Aït-Sahalia and Saglam's (2014) conjecture that fast traders are better at handling inventory risk.

Second, although we emphasize that relative latency is a key success factor for HFT firms, we also find evidence that latency is not the only dimension on which HFT firms compete. There is a small group of firms that operate at cutting edge to compete for the bulk of the trading opportunities, but there is also a large group of HFT firms that are fast but not the fastest, and they also earn positive revenues and returns, albeit lower than those of the fast ones. Thus, HFT firms presumably engage in a wider array of strategies than what the stylized theories of relative latency competition imply (Biais et al. (2015), Budish et al. (2015)). For example, Dugast and Foucault (2018) show that there is a theoretical trade-off between speed and accuracy in the interpretation of news. Traders that are not on the cutting edge in terms of speed may instead specialize in accuracy.

VIII. Conclusion

We study the role of latency in the performance of HFT firms. We document a number of statistics consistent with superior investment performance by HFT firms. There are large cross-sectional differences in performance in the HFT industry, with trading revenues disproportionally accumulating to a few firms. The fastest firms tend to earn the largest trading revenues. Although latency decreases substantially over the 5-year sample, we show that it is relative latency, not nominal latency, that helps explain differences in performance across HFT firms.

Furthermore, we examine how speed may be used in specific HFT strategies. We find evidence that relative latency is important for success in trading on short-lived information, for risk management in liquidity provision, and for cross-market arbitrage.

Finally, we explore the implications of competition on relative latency regarding HFT concentration. If small differences in latency are important, then the HFT industry should be characterized by persistence in performance, high and nondeclining concentration, and difficulty of new entry. We find evidence that is consistent with these predictions. Firm trading performance is persistent, trading revenues are high and nondeclining, as is HFT concentration, and new HFT entrants tend to be slower, underperform, and more likely to exit. Nevertheless, despite concerns in the theoretical literature about relative latency competition, we find that the cost of the HFT industry incurred to other investors is not higher than typical exchange-taker fees and is an order of magnitude lower than the effective bid–ask spread.

Supplementary Material

Supplementary Material for this article is available at https://doi.org/10.1017/S0022109018001096.

References

Aït-Sahalia, Y., and M. Saglam. "High-Frequency Traders: Taking Advantage of Speed." Working Paper, National Bureau of Economic Research (2014).

Biais, B.; T. Foucault; and S. Moinas. "Equilibrium Fast Trading." *Journal of Financial Economics*, 116 (2015), 292–313.

Boehmer, E.; D. Li; and G. Saar. "The Competitive Landscape of High-Frequency Trading Firms." *Review of Financial Studies*, 31 (2018), 2227–2276.

Bongaerts, D., and M. Van Achter. "High-Frequency Trading and Market Stability." Working Paper, available at https://www.ssrn.com/abstract=2698702 (2016).

Breckenfelder, J. "Competition between High-Frequency Traders, and Market Quality." MPRA Working Paper (2013).

Brogaard, J., and C. Garriott. "High-Frequency Trading Competition." *Journal of Financial and Quantitative Analysis*, forthcoming (2019).

Brogaard, J.; B. Hagströmer; L. Nordén; and R. Riordan. "Trading Fast and Slow: Colocation and Liquidity." *Review of Financial Studies*, 28 (2015), 3407–3443.

Brogaard, J.; T. Hendershott; and R. Riordan. "High-Frequency Trading and Price Discovery." *Review of Financial Studies*, 27 (2014), 2267–2306.

Budish, E.; P. Cramton; and J. Shim. "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response." *Quarterly Journal of Economics*, 30 (2015), 1547–1621.

Carhart, M. "On Persistence in Mutual Fund Performance." *Journal of Finance*, 52 (1997), 57–82.

Chaboud, A.; B. Chiquoine; E. Hjalmarsson; and C. Vega. "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market." *Journal of Finance*, 69 (2014), 2045–2084.

Clark-Joseph, A. "Exploratory Trading." Working Paper, Harvard University (2013).

Dugast, J., and T. Foucault. "Data Abundance and Asset Price Informativeness." *Journal of Financial Economics*, 130 (2018), 367–391.

Ellis, K.; R. Michaely; and M. O'Hara. "The Accuracy of Trade Classification Rules: Evidence from NASDAQ." *Journal of Financial and Quantitative Analysis*, 35 (2000), 529–551.

Fama, E., and K. French. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*, 33 (1993), 3–56.

Fama, E., and J. MacBeth. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy*, 81 (1973), 607–636.

Foucault, T.; J. Hombert; and I. Roşu. "News Trading and Speed." *Journal of Finance*, 71 (2016), 335–382.

Foucault, T.; R. Kozhan; and W. Tham. "Toxic Arbitrage." *Review of Financial Studies*, 30 (2017), 1053–1094.

Frank, R. "Positional Externalities Cause Large and Preventable Welfare Losses." *American Economic Review*, 95 (2005), 137–141.

Hagströmer, B., and L. Nordén. "The Diversity of High-Frequency Traders." *Journal of Financial Markets*, 16 (2013), 741–770.

Hagströmer, B.; L. Nordén; and D. Zhang. "How Aggressive Are High-Frequency Traders?" *Financial Review*, 49 (2014), 395–419.

Hasbrouck, J. "Intraday Price Formation in U.S. Equity Index Markets." *Journal of Finance*, 58 (2003), 2375–2400.

Hendershott, T., and R. Riordan. "Algorithmic Trading and the Market for Liquidity." *Journal of Financial and Quantitative Analysis*, 48 (2013), 1001–1024.

Ho, T., and H. Stoll. "The Dynamics of Dealer Markets Under Competition." *Journal of Finance*, 38 (1983), 1053–1074.

Hoffmann, P. "A Dynamic Limit Order Market with Fast and Slow Traders." *Journal of Financial Economics*, 113 (2014), 156–169.

Jovanovic, B., and A. Menkveld. "Middlemen in Limit-Order Markets." Working Paper, available at https://ssrn.com/abstract=1624329 (2015).

Kirilenko, A.; A. Kyle; M. Samadi; and T. Tuzun. "The Flash Crash: High-Frequency Trading in an Electronic Market." *Journal of Finance*, 72 (2017), 967–998.

Lee, C., and M. Ready. "Inferring Trade Direction from Intraday Data." *Journal of Finance*, 46 (1991), 733–746.

Malinova, K.; A. Park; and R. Riordan. "Do Retail Investors Suffer from High Frequency Traders?" Working Paper, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2183806 (2016).

Menkveld, A., and M. Zoican. "Need for Speed? Exchange Latency and Liquidity." *Review of Financial Studies*, 30 (2017), 1188–1228.

Petersen, M. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *Review of Financial Studies*, 22 (2009), 435–480.

Roşu, I. "Fast and Slow Informed Trading." Working Paper, available at https://ssrn.com/abstract=1859265 (2016).

Shkilko, A., and K. Sokolov. "Every Cloud Has a Silver Lining: Fast Trading, Microwave Connectivity and Trading Costs." Working Paper, Wilfrid Laurier University (2016).

Thompson, S. "Simple Formulas for Standard Errors that Cluster by Both Firm and Time." *Journal of Financial Economics*, 99 (2011), 1–10.

U. S. Securities Exchange Commission. "Concept Release on Equity Market Structure." Available at https://www.sec.gov/rules/concept/2010/34-61358.pdf (2010).

van Kervel, V., and A. Menkveld. "High-Frequency Trading around Large Institutional Orders." *Journal of Finance*, forthcoming (2019).

Van Ness, B.; R. Van Ness; and R. Warr. "The Impact of Market-Maker Concentration on Adverse Selection Costs for NASDAQ Stocks." *Journal of Financial Research*, 28 (2005), 461–485.

Weller, B. "Intermediation Chains and Specialization by Speed: Evidence from Commodity Futures Markets." Working Paper, University of Chicago (2013).

Weston, J. "Competition on the NASDAQ and the Impact of Recent Market Reforms." *Journal of Finance*, 55 (2000), 2565–2598.

Yao, C., and M. Ye. "Why Trading Speed Matters: A Tale of Queue Rationing under Price Controls." *Review of Financial Studies*, 31 (2018), 2157–2183.

Yueshen, B. "Queuing Uncertainty in Limit Order Market." Working Paper, available at https://ssrn.com/abstract=2336122 (2014).