Risk and Return in High-Frequency Trading

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

First Draft: August 2012

Current Draft: December 2016

Abstract

We study performance and competition among high-frequency traders (HFTs). We construct measures of latency and find that differences in *relative* latency account for large differences in HFTs' trading performance. HFTs 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 a variety of strategies including market making and cross-market arbitrage. We explore implications of competition on relative latency and find support for various theoretical predictions.

* Contact: Matthew Baron, Johnson Graduate School of Management, Cornell University, e-mail: baron@cornell.edu; Jonathan Brogaard, Foster School of Business, University of Washington, e-mail: brogaard@uw.edu; Björn Hagströmer, Stockholm Business School, Stockholm University, email: bjorn.hagstromer@sbs.su.se; and Andrei Kirilenko, Brevan Howard Centre for Financial Analysis, Imperial College Business School, e-mail: a.kirilenko@imperial.ac.uk

The authors would like to thank Hank Bessembinder, Tarun Chordia, Thierry Foucault, Charles Jones, Terry Hendershott, Andrew Karolyi, Robert Korajczyk, Ananth Madhavan, Katya Malinova, Maureen O'Hara, Neil Pearson, Ryan Riordan, Gideon Saar, Ronnie Sadka, Ingrid Werner, and Wei Xiong for their valuable feedback. We are grateful to Finansinpektionen for making data available for the paper. Björn 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.

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, Kozhan and Tham (2015) and Foucault, Hombert and Roşu (2016) show how competition based on relative latency can reduce market quality by increasing adverse selection of non-HFT firms. Biais, Foucault and Moinas (2015) and Budish, Cramton and Shim. (2015) show in theory how it can lead to market concentration and inefficient over-investment in speed. In these models, the fastest HFT firm responds first to profitable trading opportunities, capturing all the gains, while slower participants arrive marginally too late to trading opportunities to compete. As a result, small differences in trading speed are associated with large differences in trading revenues across firms. HFT revenues do not fall over time: regardless of how fast the market as a whole becomes, there is always at least one HFT firm with a relative speed advantage that can adversely select other traders.

Motivated by the view that competition based on relative latency differs from competition of traditional market intermediaries, this paper 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.

While HFTs 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) finds that not all HFTs choose faster colocation technology when offered, and we similarly find only about half of HFTs react to market events at time scales near the latency frontier. This suggests that many HFTs use computational power for other reasons, perhaps to better aggregate information from news feeds or order flow, and may not compete to be fastest. Despite these alternative possibilities, we find that the HFTs who are the fastest have better trading performance. We also find evidence supporting the importance of *relative* latency for trading performance and explore some of the predictions regarding market concentration.

The theoretical literature has put forward a variety of channels through which HFTs may translate speed into profitability. For example, HFTs can use speed to enhance risk management, by avoiding adverse

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

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 impact when trading with market orders, suggesting they are able to be the first to react to new information. Looking at cross-market arbitrage, we also observe the fastest firms being more responsive to information on other exchanges. Thus, there is no one single dimension through which speed is beneficial.

Our analysis uses proprietary transaction-level data with trader identifiers provided by the Swedish financial supervisory authority, Finansinspektionen. The data contain all trades of Swedish equities from January 2010 to December 2014 from all venues including regulated exchanges, multilateral trading facilities (MTFs), 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 whole picture of trading. In addition, the five-year length of our data is important in allowing us to 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, as Hagströmer and Nordén (2013) show that HFT activity is mainly concentrated in these stocks. We classify high-frequency traders as those firms that self-describe as such through their membership in the *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 website, undertakes low-latency proprietary trading.[2] The 16 firms that we identify as HFTs all have international trading operations and none of them are headquartered in Sweden. Thus, it is unlikely that the findings reported in this paper are specific to the Swedish context.[3]

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.[4] We also include risk-adjusted performance measures, including

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

[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 HFTs are highly active in this market, constituting around 30% of the trading volume and more than 80% of the order volume. Other empirical studies on this market are Breckenfelder (2013); Brogaard et al. (2015); Hagströmer, Nordén and Zhang (2014); van Kervel and Menkveld (2015); and Menkveld and Zoican (2015).

[4] Since our data set does not convey trading fees or other HFT operational costs, we are unable to directly calculate trading profits. However, in Section III, we analyze 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

returns, factor model alphas, and Sharpe ratios. We find that HFTs 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 (its 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 one 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.[5] As an example of a strategy our measure may capture, Clark-Joseph (2012) shows that HFTs use the execution of small test orders as a signal to trade on incoming order flow ahead of public order book feeds. Over our five-year sample period we show that the latency of the fastest HFTs fall substantially.

We find that relative latency, not nominal latency, drives differences in performance across HFTs. Relative latency measures how fast a HFT firm is relative to other HFTs and is captured by ranking HFTs by their calculated latency measure. Nominal latency measures how fast a HFT firm is in absolute terms and is captured by the log of a HFT firm's calculated *Decision Latency*. Our evidence is consistent with Biais et al. (2015) and Budish et al. (2015), who argue that competition on relative latency can lead to an inefficient and costly arms race. The discontinuous difference in payoffs provides strong incentives to become marginally faster than other HFTs through greater technological investment. Such competition on relative latency gives rise to a "positional externality" (Frank, 2005), since a firm that becomes faster increases the relative latency of its competitors, which can in turn lead to an inefficient over-investment in speed.

We find that firms that are among the five fastest HFTs, and in particular the fastest single HFT firm, earn substantially higher revenues than other HFTs. It is not being fast that allows an HFT firm to capture trading opportunities, it is being faster than others, consistent with Biais et al. (2015) and Budish et al. (2015). Furthermore, we find that the fastest HFTs capture more trading opportunities and have higher

allow 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.

[5] 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 IV.C.

risk-adjusted revenues, but they do not earn higher revenue margins per SEK traded. The differential finding suggests that on a per-trade basis, the fastest HFTs are no more accurate than other traders at processing and analyzing information (trade quality), but their latency advantage allows them to capture more trading opportunities (trade quantity) without taking on higher risk.[6]

As a robustness check, we construct various alternative approaches to measuring HFT latency. For example, *Queuing Latency*, captures the race to be at the top of the order book, as motivated, for example, by theoretical work by Yueshen (2014) and empirical findings by Yao and Ye (2015). 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. Importantly, *Queuing Latency* does not rely on time stamps, making it robust to potential time stamp noise, and it is potentially better at capturing latency of HFTs that do not use market orders. Nevertheless, we find qualitatively similar results as with our primary latency measure.

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 March 14, 2011 and the "10G Colocation" upgrade first offered on September 17, 2012. These colocation upgrades lead some, but not all, HFTs to get faster.[7] We compare the change in trading performance for HFTs 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. Some theories view fast traders as using speed to trade on short-lived information, whether in reaction to news, order flow, or latency arbitrage (Cartea and Penalva 2012; Foucault et al., 2015; Foucault et al., 2016; Biais et al., 2015; and Roşu 2015). Other theories view speed as a way to avoid adverse selection and inventory costs (Jovanovic and Menkveld, 2015; Aït-Sahalia and Saglam, 2014; Hoffmann, 2014). Menkveld and Zoican (2015) posit that both these types can co-exist in equilibrium.

We examine the role of relative latency in both channels, both in the general setting as well as in a specific cross-market strategy. We proxy the short-lived information channel by the ability of a market order to predict price changes over the next 10 seconds, and the risk management channel by the ability of a passive order to capture a large realized spread. Relative latency is associated with better performance

[6] We use aggregate revenues as the main measure of performance to capture trade quantity. If strategies are not easily scalable, trade quality measures such as per-trade revenues are less relevant for comparing firms (Chen, Hong, Huang, and Kubik, 2004). A firm that has high revenues per trade but that captures few trading opportunities may not be considered a strong performer.

[7] This differential effect on HFTs may be explained by the fact that not all HFTs immediately subscribe to the colocation upgrade, as documented by Brogaard et al., 2015; and even that among those that do, not all HFTs may be equally able to translate this technology into faster trading.

through both channels. As a specific strategy, we study cross-market arbitrage by examining HFTs' equity trading following changes in the price of index futures. In the second after a change in the index futures price, the fastest HFTs are more likely than other HFTs to aggressively trade in individual equities in the direction of the futures price change. The fastest HFTs 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.

Finally, we explore predictions regarding the effects of relative latency on market concentration. If the traditional view of market-making competition holds (Ho and Stoll, 1983; Weston 2000), we expect the alpha generated by HFTs and the concentration of revenues to disappear as the industry matures. Alternatively, if HFTs 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 adversely select other traders. Additionally, rents remain concentrated among the fastest HFTs, as slower HFTs arrive marginally too late to trading opportunities to compete.

Consistent with the predictions of the effects of competition on relative latency, we find that the HFT industry is concentrated among a few firms. In contrast to the traditional view that increased competition over time leads to lower profits, HFT concentration of trading revenues and trading volumes are high and non-declining over the five year sample, despite new HFT firm entry and a decline in overall HFT latency. We furthermore find that new HFT entrants are typically slower, earn lower trading revenues, and are more likely to exit, which likely reinforces concentration in the HFT industry.[8]

II. Data and Methodology

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

[8] A previous version of this paper analyzes HFT performance in the E-mini S&P 500 futures contract over a two year period from 2010 to 2012. While 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 paper are replicable, have external validity, and are robust to differences in market structure.

supervision of one of the national financial supervisory authorities must report all their transactions with financial instruments to TRS.

The TRS data has two features that make it highly suitable for the analysis of revenues in equity trading. First, the scope of the reporting obligation spans transaction at all trading venues, including regulated exchanges, MTFs, and dark pools. This is important given the high degree of fragmentation of volume in European equity trading. Second, TRS contains 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 trader identifiers are necessary to identify HFTs and to analyze revenues in the cross-section of firms. Finally, the TRS data contains standard transaction-level variables such as date, time, venue, price, currency, quantity, and a buy/sell indicator. See Appendix Section A1 for information about the filtering procedures applied to the TRS data.[9]

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 HFTs 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.[10] 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 January 4, 2010 to December 30, 2014.

We match the TRS transactions to transaction-level data available from the *Thomson Reuters Tick History* (*TRTH*) database. The TRTH database is accessed through the *Securities Research Centre of Asia-Pacific* (SIRCA). The purpose of the matching is twofold. First, whereas the TRS data has second-by-second time stamps, TRTH has time stamps at a microsecond granularity. Through the matching we can assign microsecond time stamps to the TRS data, which is important for our latency measurement. Second, TRTH also contains order book information recorded on a microsecond frequency synchronized to the transaction data. This enables us to assess the status of the order book just before each TRS transaction,

[9] 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.

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

which is necessary to measure, for instance, the effective spread and to determine whether the trade was initiated by the buyer or the seller, following Lee and Ready (1991).[11]

B. Trading Venues and Stock Characteristics

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 by trading volume in our data are Chi-X, BATS, Turquoise (all based around London) and Burgundy (based in Stockholm). In February 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.

Table 1 reports descriptive statistics for the sample stocks. *Market Capitalization* at closing prices on December 31, 2014 ranges from 13,877 million SEK (henceforth MSEK) for *SSABa* to 475,595 MSEK for *HMb*, the equivalent of 1.78 to 60.91 billion USD, converted at the exchange rate of December 31, 2014. In the U.S. equities market, these stocks would be labeled as large or mid-cap stocks.

INSERT TABLE 1 ABOUT HERE

Daily Trading Volume refers to trading at NASDAQ OMX Stockholm only and is reported in MSEK. *Daily Turnover* is the Daily Trading Volume divided by Market Capitalization, expressed in percentage points. *Tick Size* is the average minimum price change. *Quoted Spread* is the average bid-ask spread prevailing just before each trade; and *Effective Spread* is the trade value-weighted average absolute difference between the trade price and the bid-ask midpoint. All spread measures are based on continuous trading at NASDAQ OMX Stockholm, expressed relative to the bid-ask spread midpoint, and presented in basis points. The *Tick Size* and the *Quoted Spread* are halved to be comparable to the *Effective Spread.* The *Daily Turnover* across stocks is 0.60% and the *Quoted Spread* and *Effective Spread* vary between 2 and 6 bps. The more liquid stocks in our sample have a turnover and spread similar to the US large-cap stocks studied by Brogaard, Hendershott, and Riordan (2014). The *Tick Size* for many stocks is close to the *Quoted Spread*, indicating that market tightness is frequently bounded by the tick size.

[11] Concerns about the accuracy of the Lee-Ready algorithm (see Ellis, Michaely and O'Hara, 2000) have limited applicability in this data set. First, trades inside the quotes are uncommon. This is due to 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. For each trade recorded in TRTH, there is also a quote update (usually with the same microsecond time stamp) reflecting how the trade influences the order book.

Finally, we report *Volatility*, the average 10-second squared basis point returns, calculated from bid-ask midpoints; and an index for the degree of volume fragmentation. The *Fragmentation Index* is defined as the inverse of a Herfindahl index of trading volumes across the five largest trading venues (BATS, Burgundy, Chi-X, NASDAQ OMX Stockholm, and Turquoise). The procedure implies that fragmentation is measured on a scale from one to the number of trading venues considered, which in our case is five.[12] *Volatility* ranges from 3 to 17 squared basis points, and the *Fragmentation Index* varies across stocks between 1.76 and 2.32.

C. HFT Identification

Previous studies classify HFTs according to observed trading behavior (as in Kirilenko, Kyle, Samadi and Tuzun, 2015) or using an exchange-defined classification (Brogaard et al., 2014). We define HFTs as those who self-describe as HFTs by including firms that are members of the *FIA-EPTA* or that according to their own website 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 HFTs have the characteristics usually associated with them: high trading volume, short investment horizons, and tight inventory management (Securities and Exchange Commission, 2010).

To include an HFT firm, we also require it to trade at least 10 MSEK a day, about 1.05 million USD at the exchange rate on December 31, 2014, for at least 50 trading days of the 1,255 trading days in the sample. We find 25 firms who self-describe as HFTs, 16 of which satisfy the volume criteria and form our sample of HFTs. 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 HFTs, but that do not satisfy the volume criteria together represent only 0.13% of the total HFT trading volume, and 0.85% of the firm-day observations.[13]

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 that

[12] If there are *N* trading venues and they all have equal shares of the trading volume, the index takes its maximum value *N*. If all trading volume is concentrated to one venue the index takes its minimum value, which is 1. The index design is similar to the Fidessa Fragmentation Index, more details of which can be found at http://fragmentation.fidessa.com/faq/#faq2

[13] Due to confidentiality requirements, we cannot report the full list of names of the 25 HFTs covered in the proprietary data set. However, in Appendix Section A2, we use public trading records to report the names of 19 HFTs who trade at NASDAQ OMX Stockholm as members. The HFTs not listed in Appendix Section A2 therefore trade only at other trading venues or as clients of other members at NASDAQ OMX Stockholm.

are ex-ante expected to be profitable to the HFT firm, 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 capture 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. 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*.

We assume zero beginning-of-day inventory positions as a way to overcome potential data errors. Even minor errors in inventory can accumulate over time, leading to large and persistent (unit root) errors if left uncorrected. Therefore, we zero beginning-of-day inventories so that any potential errors do not affect more than one day. This assumption is relatively innocuous because we show below that most HFTs usually end the day near a zero position anyway (see Table 2). In the Appendix Section A3, we compare our main method of calculating trading revenues with three alternative approaches, one of which relaxes the assumption of zero inventory at the start of the trading day and cumulates daily net inventory positions over the full sample. We conclude that alternative definitions of *Revenues* yield similar results.

To capture risk-adjusted performance we measure *Returns*, factor model *Alphas* (one, three, or four factors), and the *Sharpe Ratio*. Through the use of risk-adjusted performance measures, we assess whether HFTs 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. Ait-Sahalia and Saglam (2014) show that fast market-makers are better at handle inventory risk, and Hoffman (2014) shows that fast traders are able to avoid adverse selection risk. In its IPO prospectus, Virtu, an HFT firm in our sample, states: *"As a result of our real-time risk management strategy and technology, we had only one losing trading day during … a total of 1,238 trading days."* [14]

Returns are calculated by dividing *Revenues* of 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 five-year sample. HFTs' inventories generally exhibit sharp, well-defined

[14] The prospectus is available at
https://www.sec.gov/Archives/edgar/data/1592386/000104746914002070/a2218589zs-1.htm

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 equities markets.[15] *Returns* can thus be viewed as the performance achieved relative the capital allocated to the trading operation. *Returns* are calculated at daily frequencies but throughout the paper are reported in annualized terms.

Factor model *Alphas* are computed for each HFT firm over the entire sample using the standard Fama-French model (Fama and French, 1993) and the Carhart (1997) momentum factor. 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 website, using the full sample of Swedish stocks traded on NASDAQ OMX Stockholm. Methodological details concerning the construction of these factors and validation exercises can be found in the Appendix Section A4.

The annualized *Sharpe Ratio* for each HFT firm is calculated using daily observations as $\frac{\mu_i - r_f}{\sigma_i} * \sqrt{252}$, 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. Whereas *Returns* and factor model *Alphas* rely on the assumption that market capitalization can be proxied by the maximum inventory position of the trading firm, the *Sharpe Ratio* does not. To see this, note that if the risk-free rate can be neglected as it is nearly zero for much of the sample period, the *Sharpe Ratio* is equivalent to: $\frac{\mu(Revenues)_i}{\sigma(Revenues)_i} * \sqrt{252}$. The equity capitalization is therefore irrelevant for calculating the *Sharpe Ratio*.

We capture quality performance with the *Revenues per MSEK Traded* measure. The quality dimension of performance measures the ability to enter trades with a high revenue margin. *Revenues per MSEK Traded* is calculated daily as *Revenues* divided by *Trading Volume*.

The performance measures do not account for trading fees and liquidity rebates. We show in Section III.C that an adjustment for estimated exchange fees and liquidity rebates does not change the conclusions of the paper.

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 the response as a subsequent aggressive execution by the same firm. Examples of why HFTs would attempt to

[15] In Section III, we show that HFT returns calculated this way are comparable in magnitude to those from 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.

trade aggressively immediately after a passive execution include test orders described by Clark-Joseph (2012) and "scratch" trades described by Kirilenko et al. (2015). The HFT firm cannot control the timing of the passive trade but 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 one second. The time-stamp difference between the two trades in each case forms an empirical distribution of response times. 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. [16, 17]

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.

By excluding cases where the two trades are recorded at different trading venues we avoid potential problems related to that time-stamps are not perfectly synchronized across venues. Additionally, we can test whether within-market *Decision Latency* also explains success at arbitrage across markets (see Section V).

There are numerous signals that may trigger HFTs to react swiftly, including news events, order book gaps, and block orders. The inherent problem of signal-to-response latency measures is that HFTs employ different strategies and put different weights to different signals. We argue that it is likely that HFTs

[16] 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 sub-millisecond 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 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 HFTs 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.

[17] 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 IV.C.

respond to signals affecting their own portfolio such as a passive execution. Our measure of *Decision Latency*, while not perfect, captures an important dimension of latency that varies across market participants. While 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 HFTs that do not tend to follow passive executions with immediate aggressive executions.[18] However, these limitations should result in underestimating, not exacerbating, the role of speed in performance. Furthermore, in Section IV.C we show that our results are robust to two alternative measures of latency, *Queuing Latency* and *Mean Latency*.

Figure 1 plots *Decision Latency* over the sample period 2010-2014. HFTs are grouped by their relative rank of latency per month; the categories are Top 1, Top 1-5, and all HFTs. Over the sample period latency decreases for HFTs in the top 5: for example, the latency of the Top 1 HFTs decreases from around 62 microseconds in 2010 to around 10 microseconds in 2014. The relative reduction in latency is much greater for Top 1-5 HFTs, who start out in 2010 with latencies of over 1,280 microseconds and converge in latency to the Top 1 HFT by 2014. In contrast, *All HFTs*, which disproportionately picks up the slower HFTs, remains relatively constant with an average latency of 25 milliseconds over the entire sample period. The finding that HFTs outside the Top 5 do not achieve lower latencies over time is consistent with the finding of Brogaard et al. (2015) that not all HFTs choose to be the fastest when given the opportunity to choose a faster colocation technology.

INSERT FIGURE 1 ABOUT HERE

The magnitude of latency recorded for the fastest HFTs in this paper is consistent with statements about the INET trading system used at NASDAQ OMX Stockholm. In marketing materials from 2012, NASDAQ states that their trading system delivers "sub-40 microsecond latency."[19] At that time, our fastest measured latency is around 60 microseconds.[20]

[18] *Decision Latency* cannot be measured for HFTs that trade exclusively using either aggressive or passive orders. In our sample, 2.2% of the firm-months are subject to this limitation, but 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.

[19] http://www.nasdaqomx.com/digitalAssets/82/82655_markettechoverview_oct2012.pdf

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

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 March 2011 and September 2012. The fact that *Decision Latency* decreases following the technology upgrades provide suggestive evidence that our latency measure indeed captures reaction time. While it is difficult to assess the impact of the 2010 INET upgrade since it comes at the start of the sample, the colocation upgrade of 2012 is followed by a decline in latency for the top 5 HFTs. The fact that latency falls subsequent to colocation upgrades is a valuable validation of our measure. In Section IV, we use the 2011 and 2012 colocation upgrades to provide evidence on a causal relation between relative latency and trading performance.

III. Characterizing HFT Performance

A. HFT Performance in the Cross-Section

We document the risk and return characteristics of individual HFT firms. In Table 2 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.

INSERT TABLE 2 ABOUT HERE

The median HFT firm realizes an average daily *Revenues* of 6,990 SEK, or 56.5 SEK *Revenues per MSEK Traded*. It has a daily *Trading Volume* of 64 MSEK, an annualized *Sharpe Ratio* of 1.61, and a four-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.[21]

We find considerable performance variation in the cross-section of HFTs. The cross-sectional distributions are skewed towards a few high performers. For example, firms in the top 90th percentile generate *Revenues* of 61,354 SEK per day, compared with 6,990 for the median; a *Sharpe Ratio* of 11.1, compared with 1.61 at the median; *Revenues per MSEK Traded* of 472.2, compared with 56.5 at the median; and a four-factor annualized *Alpha* of 89%, compared with 9% at the median.

[21] Appendix Section A5 analyzes HFT performance after accounting for potential maker-taker fees and liquidity rebates. Even after accounting for the most conservative possible fees and/or rebates, 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 HFTs at the top strongly outperforming their competitors.

HFTs are diverse in terms of other trading characteristics, too. Beyond performance, 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 HFTs in the Securities and Exchange Commission's Concept Release on Equity Market Structure (2010) and with functional-based approaches for HFT classification (Kirilenko et al., 2015), most, though not all, HFTs tend to have low intraday and end-of-day inventories. HFTs vary in their *Aggressiveness Ratio*, with some nearly all active or passive and others mixed. The average aggressive ratio is 53%. Consistent with Figure 1 and the discussion above, there is also substantial variation in *Decision Latency* across HFTs, from 42 microsecond latency at the 10th percentile to 0.5 second latency at the 90th percentile, a finding we explore in Section IV. Notably, the 0.5 second latency for some HFTs to process information and react with a market order is slow for automated traders but still fast relative to human reaction time.

B. Comparison of Trading Revenues to Trading Profits Based on Public Filings

The data do not convey trading fees or other operational costs and so we are unable to directly calculate trading profits. However, regulatory filings of five major HFT firms (Virtu, 2011-2015; Knight Capital Group, 2013-2015; GETCO, 2009-2012; Flow Traders, 2012-2015; and Jump, 2010) allow a comparison of trading revenues and trading profits. A potential concern in our analysis of HFT performance is that firms with higher trading revenues may have higher fixed costs. That is, firms with higher trading revenue may also incur higher costs to produce better performance. If true, then trading revenues may not be a good proxy for firm profitability. We show that this is not likely the case.

Table 3 reports *trading revenue*, *trading costs*, *trading profit margins*, and *trading returns* calculated from annual reports, IPO prospectuses, and SEC disclosures for five HFT firms for which public data is available.[22] Trading costs are broken down into several categories such as trading and clearing fees, data costs, financing costs, equipment and technical costs, all expressed as a percent of trading revenues. Trading costs also include *depreciation and amortization*. This serves as a control for investments that a firm may have undertaken in years preceding the public data coverage.

[22] Jump Trading was never a public company like the other four but nevertheless filed publicly available SEC disclosures containing trading revenues and profits for 2010 (see, https://www.bloomberg.com/news/articles/2014-07-23/don-t-tell-anybody-about-this-story-on-hft-power-jump-trading).

INSERT TABLE 3 ABOUT HERE

We make two observations. First, trading profit margins are high, ranging between 27.4% and 64.5% of trading revenue for all four firms. Approximately 40-80% of the HFT costs are per-trade fees: brokerage fees, exchange and clearance fees, and financing costs. The fixed (i.e. not per-trade) costs, including communications and data processing, equipment, administrative and technology costs, make up only 15-30% of the total costs. As a result, we conclude that fixed costs, which include costs related to technological investment and colocation services, are small relative to trading revenues, making it unlikely that firms with the highest trading revenues face higher investments costs that would substantially reduce their net profits.

Second, as a percentage of trading revenues, the fixed costs do not vary substantially across firms, suggesting that revenues are not correlated with fixed costs in percentage terms. For example, in 2014, KCG had double the trading revenue of Virtu and five times the trading revenue of Flow Traders, but the total fixed costs as a percentage of trading revenue show no pattern (22.7% for KCG; 17.7% for Virtu; 27.2% for Flow Traders). There is also no clear time trend in fixed costs within each firm to suggest that higher trading revenues periods might be correlated with higher fixed costs. All else being equal, the stability of the fixed costs suggests that firms with higher trading revenues also have higher profits. As such, HFT revenue variation is likely a close proxy for variation in HFT profits.

Table 3 reports *trading returns. Trading returns* are calculated in two ways based on different capitalization measures: *trading revenue / (trading assets minus trading liabilities)* and *(trading revenue / book equity)*. From these public filings in which capitalization is directly observable, we find trading returns to range from 60% to almost 237%, depending on the firm. This suggests the returns computed in Section III.A are of a reasonable magnitude.

IV. The Role of Speed in Performance

Having documented the performance of HFTs, we now test our main hypothesis about speed and HFT trading revenues. While most theories in which HFTs earn profits posit that fast traders should have an advantage, other theories suggest that traders of different speed can specialize along other dimensions (Weller, 2013; Roşu, 2015). According to these models, a relatively slow market intermediary could compensate by 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):

$$Performance_{i,t} = \alpha_t + \beta_1 \log(Decision\ Latency)_{i,t} + \beta_2 \mathbf{1}_{\text{top 1 } i,t} + \beta_3 \mathbf{1}_{\text{top 1-5 } i,t} + \gamma' controls_{i,t} + \text{month-FEs} + \epsilon_{i,t}, \quad (1)$$

where $Performance_{i,t}$ 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 Eqn. (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.

The independent variables $\mathbf{1}_{\text{top 1 } i,t}$ and $\mathbf{1}_{\text{top 1-5 } i,t}$, are indicators for whether a given firm is ranked in the Top 1 or Top 1-5 by speed in a given month. Nominal latency is captured using *Decision Latency*, while the $\mathbf{1}_{\text{top 1 } i,t}$ and $\mathbf{1}_{\text{top 1-5 } i,t}$ indicators capture relative speed. Since *Decision Latency* can vary widely across firms, from the microsecond to the second level (see Table 2), the relationship between trading speed and trading revenues is best captured by taking logs. The indicator variables capture the potentially non-linear 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 performance, including measures of their risk-bearing capacity and strategies. These variables include the *End-of-Day Inventory Ratio*, *Max. Intraday Inventory Ratio*, *Investment Horizon*, and the *Aggressiveness Ratio*, which are defined in Section III.A and calculated on the monthly frequency for each HFT firm. *Max. Intraday Inventory* is used in the denominator to calculate *Returns* and is thus omitted when *Returns* is the dependent variable.

The continuous independent variables, namely, log(*Decision Latency*) and the control variables, are normalized to be in units of standard deviations. Month fixed-effects absorb time-varying market conditions, including market trading volume and volatility. Following Petersen (2009) and Thompson

(2011), standard errors are dually clustered by firm and month to account for correlations both across firms and over time. Table 4 reports coefficient estimates for various specifications of the model described in Eq. (1).

INSERT TABLE 4 ABOUT HERE

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 $\mathbf{1}_{\text{top 1 } i,t}$ and $\mathbf{1}_{\text{top 1-5 } i,t}$ indicators of relative latency 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 log(*Decision Latency*) suggest that relative speed matters more than nominal speed. Specifically, the estimates in Column 3 show that being among the five fastest HFTs (Top 1-5) predicts average daily trading revenues that are SEK 15,451 higher than the for the HFTs outside the top five. Being the fastest (Top 1) provides on average daily trading revenues of SEK 24,639 in addition to the revenues from being in the Top 1-5: the coefficient on the *Top 1* dummy tests the difference in revenues between being Top 1 and Top 1-5, which is found to be statistically significant.[23] HFTs that are not among the five fastest in a given month have average daily revenues reflected by the intercept of SEK 10,894. These numbers imply that the revenues of the fastest HFT firm are around five times higher than those of the "slow" (non-Top 1-5) HFTs.

Several of the control variables are related to trading revenues. For example, a one standard deviation increase in the *Max. Intraday Inventory* is associated with decreased daily *Revenues* of 21,008 SEK, suggesting that HFTs that have tighter inventory management perform better. Similarly, HFTs 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 HFT firm size does not drive the relation between *Revenues* and relative latency. Furthermore, we find that the *Sharpe Ratio* is higher for HFT firms

[23] Appendix A6 repeats the analysis presented here but breaks down the *Top 1-5* dummy variables into individual dummy variables for the fastest HFTs: *Top 1*, *Top 2*, *Top 3*, *Top 4*, and *Top 5*. It shows that, even among the top five HFTs, the faster firm tends to perform better and that performance is monotonic in relative latency.

with lower relative latency. This demonstrates that the relation between *Revenues* and relative latency is not driven by the risk of the trading strategies applied.

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 HFTs use latency advantages to better obtain and aggregate information in order to predict future price changes we would expect the fastest HFTs to have the highest revenues per trade. However, according to Table 4, 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 faster HFTs are able to capture a larger trading quantity, but that trading speed is not an important determinant of trading quality. It appears that the fastest HFTs 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.

The results presented in this section show that HFTs with relatively low latency are more successful in capturing trading opportunities and earning higher revenues. To alleviate concerns about a spurious relationship between speed and performance, in the next section we focus on a quasi-experimental setting where the trading speed changes for some firms but not for others.

B. Evidence from two colocation upgrades

To address the potential endogeneity concern that another variable correlated with HFT latency might be instead driving trading performance, we put forward evidence from a quasi-experimental setting studying two colocation upgrades that cause some HFTs to increase their relative speed.

On March 14, 2011 and September 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 HFTs improving their latency rank, and find evidence in support of a causal relation between relative latency and trading performance.

The September 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 paper. In particular, they find that only about half of the affected members immediately subscribed to the new connection type. Likely as a result, we find that some, but not all, HFTs improve their latency shortly after the upgrade becomes available.

Specifically, we measure *Decision Latency* before and after the event, and compare the change in trading performance for HFTs that become *relatively* faster through the colocation upgrade to HFTs 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 HFTs 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 March 14, 2011 event and one HFT firm for the September 17, 2012 that improve their latency rank and refer to them as *Faster*. We compare that group to three HFT firmss for the March 14, 2011 event and one HFT firm for the September 17, 2012 event that decline in latency rank, which we refer to as *Slower*. All other HFTs who have unchanged relative latency – including some who get nominally, but not relatively, faster – are excluded from this analysis.

Table 5 reports the trading performance measures for *Faster* and *Slower* HFTs, before and after the colocation upgrade event. The results suggest that the group of HFTs 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 HFTs 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 HFTs capture fewer trading opportunities. As seen by the difference-in-difference estimate in the bottom line of Table 5, 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 six degrees of freedom.

INSERT TABLE 5 ABOUT HERE

The difference-in-difference estimates are large, positive, and often statistically significant due to the consistency and magnitude of the change, despite the 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 5 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 HFTs that do not mix order types. Second, the microsecond time stamps reported by 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, though 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*, and re-estimate the results of Table 4, which analyzes the connection between latency and various measures of HFT performance. The results are qualitatively unchanged.

To mitigate the time stamp noise, we measure *Mean Latency*. We use the same distribution of latencies as for *Decision Latency*, but instead of calculating the 0.1% quantile, we define *Mean Latency* as the mean of all latencies that are shorter than one millisecond. By using a central moment rather than an extreme quantile, we expect the time-stamp noise to cancel out, relying on the central limit theorem. This approach comes at the cost of not picking the cases where HFTs operate at their very fastest speed.

Our second alternative latency measure, *Queuing Latency*, circumvents both concerns described above. For this measure, we exploit price changes that lead to a gap in the limit order book. As modelled by Yueshen (2014), if the price change is viewed as temporary, fast traders rush in to capture the top-of-queue limit order position in the emerging gap. When the price changes and a new tick opens up, *Queuing Latency* measures how often each HFT firm submits the first limit order and thus gets to the top of the queue. A higher value corresponds to lower latency. Note that this measure does not use time stamps and, furthermore, that simply more trading or limit order submissions does not help one get to the top of the

queue: given the brief window when a new tick opens up, the chances of a randomly submitted order ending up first is negligible.

The measurement procedure for *Queuing Latency* involves the following three steps. First, we identify trades that consumes all available liquidity at a price level (*gap-opening trades*). Second, for each gap-opening trade we identify the next trade at the same price level as the *gap-filling trade*. We retain the passive counterparty of all gap-filling trades that: (i) are in the same direction as the corresponding gap-opening trades; (ii) occur within 10 seconds after the corresponding gap-opening trades; (iii) do not have the same broker as buyer and seller; (iv) do not have the same passive counterparty as the corresponding gap-opening trades.[24]

We repeat the analysis in Table 4 for the two alternative measures of latency and report the findings in Table 6. As in Section IV.A, we use the alternative latency measures to rank HFTs by latency in each month and construct new *Top 1* and *Top 1-5* rank dummies. Note that in Panel A, which looks at *Queuing Latency*, we represent nominal latency with log(*Queuing Latency + 1)*, so that we do not take the log of zero, which is the lowest possible value that *Queuing Latency* can attain.

INSERT TABLE 6 ABOUT HERE

Panel A reports the results for *Queuing Latency*. In Column 1, there is an association between nominal speed and trading performance. In contrast to Table 4, we expect a positive coefficient, since lower latency is represented by a higher value of *Queuing Latency*. In Columns 2 and 3, when the *Top 1* and *Top 1-5* rank dummies are added, the coefficient on nominal latency now becomes insignificantly different from zero, while the magnitudes of the estimates on the *Top 1* and *Top 1-5* rank dummies are large and significant. Thus, as before, relative latency is more important than nominal latency. Similar results can be seen for *Returns*, the *Sharpe Ratio*, and *Trading Volume*. However, as usual for the quality measure, *Revenues per MSEK Traded*, the results are small in magnitude and not significant.

Panel B reports the results for *Mean Latency* and tells a similar story. The only main difference between Panel B and Table 4 is that the coefficient on the *Top 1-5* rank dummy is generally statistically significant, but the one on *Top 1* is not, although it is still generally positive and large in magnitude. So

[24] The reasoning behind the criteria is as follows. (i) This criterion avoids cases where the market order leading to the gap-opening trade posts its residual volume as a limit order. That is a mechanical way to cease the top-of-queue position. Execution of such limit orders is however always in the opposite direction relative the gap-opening trade. (ii) This criterion ensures that there really is a race to capture the trading opportunity. (iii) This criterion avoids the influence of the *internal* priority rule at NASDAQ OMX Stockholm. Under this rule, a broker with a limit order posted at a given price level has priority to be executed if a market order from the same broker is executed at that price. (iv) This criterion ensures that we do not capture iceberg orders that automatically converts hidden liquidity to visible if the previously visible part of the order is executed.

while being relatively fast according to this measure is still important, as reflected in the coefficient on the *Top 1-5* dummy, *Mean Latency* may not be the best measure to capture the very fastest HFTs at their peak potential, when extreme low latencies are needed to outperform competitors. Nevertheless, as a central tendency and not an extreme value, *Mean Latency* is useful for assessing the robustness of our main measure, *Decision Latency*.

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. (2015) 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 discussed above. 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 trading on short-lived information, we measure active *Price Impact* as the basis point change in the bid-ask spread midpoint from just before a trade initiated by an HFT firm to ten seconds after. To capture risk management, we measure passive *Realized Spread* as the basis point difference between the transaction price and the bid-ask spread midpoint ten 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 on a firm-stock-month frequency 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. As 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 ten seconds later.

We re-estimate Eq. (1) at the firm-stock-month level with *Price Impact* and *Realized Spread* as the dependent variables. Unlike in Table 4, we disaggregate by stock to control for stock-level characteristics, such as the *Quoted Spread* and *Tick Size*, which may affect *Price Impact* and *Realized Spread*. As before, the $\mathbf{1}_{\text{top 1 } i,t}$ and $\mathbf{1}_{\text{top 1-5 } i,t}$ are indicators that capture relative latency, whereas log(*Decision Latency*) captures nominal latency. The estimation is run with and without control variables.

The control variables include both HFT firm characteristics and stock characteristics. The HFT firm characteristics are *End-of-day Inventory*, *Max. Intraday Inventory*, *Investment Horizon*, and *Aggressiveness Ratio*, defined as in Section III.A. Since these variables are firm-characteristics, they are assigned the same value across stocks. The stock characteristics are measured on stock-month frequency and are defined as in Section II.B: *Volatility*, *Fragmentation Index, Tick Size*, and *Quoted Spread*. The *Non-HFT Trading Volume* is the daily sum of SEK trading volume in each stock that does not involve HFTs. All continuous independent variables are in units of standard deviations. Standard errors are dually clustered by firm-stock and month. The results are reported in Table 7.

INSERT TABLE 7 ABOUT HERE

As with HFT performance in general, we find for the specific channels that relative latency, not nominal latency, drives performance. The coefficients for the $\mathbf{1}_{\text{top 1-5 } i,t}$ relative latency dummy is statistically significant and economically meaningful for both the *Price Impact* and the *Realized Spread*. For example, being in the Top 1-5 by speed increases *Price Impact* by 0.645 bps, which can be related to the intercept of 3.96 bps. That means that the five fastest HFTs have 16% (0.645 / 3.96) higher price impact than other HFTs. In addition, the fastest HFT outperforms the price impact of other HFTs by another 0.34 basis points (0.34 / 3.96 = 9%). For the *Realized Spread* the Top 1-5 coefficient is 0.477 bps, whereas the intercept is insignificantly different from zero. This indicates that being among the fastest HFTs is important for being successful at the risk management required for passive trading. The results are robust to inclusion of the control variables.

We conclude that relative latency is important both for improving trading on short-lived information and for risk management. This is consistent with theoretical models, such as Foucault et al. (2016) on active trading; and 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 re-examine both channels by focusing on cross-market trading between the futures market and equities. Specifically, we test if faster HFTs are more likely than slower HFTs 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 HFTs are less likely than slower HFTs to be adversely selected in a passive trade in equities 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. (2015).

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

$$\Pr[Fast\ HFT\ Trades] = \Phi[\beta\ News + \gamma' controls + \ Stock\ FEs]. \quad (2)$$

The unit of observation is a trade. To capture who is 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. We define *Fast HFT* as being either *Top 1* or *Top 1-5* of HFTs in terms of *Decision Latency* in each month; *Slow HFT* are those not among the top five. The dependent variable is 1 when a *Fast HFT* executes an equities trade in the subsequent 1-second and 0 if a *Slow HFT* does it. Thus, our results can be interpreted as the increased probability of a *Fast HFT* trading in equities relative to a *Slow HFT*, in response to "news" in the futures market.

News is defined to be ±1 when the return on the OMXS30 futures during a one-second window preceding the stock trade is large, defined as when the absolute return exceeds the top decile among non-zero absolute returns of that month, and 0 otherwise. *News* takes the value +1 if the active party trades in 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 also control for the following variables, which may affect the probability of *Fast HFT*s doing cross-market arbitrage. *Lagged Volatility* is the average second-by-second squared return (multiplied by 1,000) over the previous ten seconds; *Lagged Volume* is the SEK trading volume (divided by 100,000) over the previous ten 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, the sample being all trades initiated through the submission of a market order by a HFT firm, are reported in Table 8, Panel A. Similarly, the estimates for passive trading, the sample being all trades where liquidity is provided by an HFT firm, are presented in Panel B. For computational tractability, regressions are run monthly for all month in 2010-2014. Similar to the Fama and Macbeth (1973) procedure, the monthly coefficients are averaged across months to produce the estimates

reported in Table 8. To assess the economic magnitudes, we report marginal effects. The marginal effects show the increased probability of *Fast HFT*s 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.

INSERT TABLE 8 ABOUT HERE

We find that the *News* coefficients are positive and statistically significant for active trading (Panel A). Based on the marginal effects, the fastest HFT firm (*Top 1*) is 0.6% more likely to actively trade in equities subsequent to "news" arrival in the futures market, relative a trader outside of the Top 5. The *Top 1-5* HFTs show similar results. Overall, we conclude that faster HFTs 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. (2015). For passive trading, the *News* coefficients are negative (Panel B), indicating that fast HFTs are less likely to get caught in a passive equities trades that incur adverse selection costs to the liquidity provider. The result is not statistically significant for the *Top 1* HFT firm, but is statistically significant at the 1% level for the *Top 1-5* HFTs. This is in line with Foucault et al. (2015), who find that the probability of toxic arbitrage is related to the latency of arbitrageurs relative the latency of liquidity providers.

VI. Implications of Speed for the HFT Industry

Sections IV and V show that relative differences in latency can help explain the variation in HFT performance. If trading speed drives performance, why are HFT firms not competing away their trading revenues, either through new entry, higher trading volumes, or increased investment in speed and trading sophistication?

According to the traditional view of market-making competition (Ho and Stoll, 1983; Weston 2000), the alpha generated by HFTs and the concentration of revenues should disappear as the industry matures. Alternatively, if HFTs compete on relative latency, then increased competition should not 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 adversely select other traders. Additionally, rents remain concentrated among the fastest HFTs, as slower HFTs arrive marginally too late to trading opportunities to compete.

We explore some of the prediction of relative latency regarding market concentration.[25] The HFT literature on competition on relative latency makes several predictions: persistence in performance, both at the firm-level and industry-wide level, high concentration of HFT revenues and trading volume, and difficulty of new entry. We examine each of these predictions and find evidence that all apply for the HFT industry.

A. Persistence in Firm-Level Performance

We first test for persistence at the firm-level. Large differences in HFT firm performance could potentially be driven by luck. For instance, in a model of identically skilled HFTs, all engaging in strategies with right-skewed distributions of performance, some will happen to outperform. Persistence shows that something other than luck drives a firm's performance.

There is an extant literature showing that performance for actively-managed mutual funds in period *t* generally does not predict performance in period *t+1* (Carhart, 1997). However, there is evidence of persistence by some investors (Jagannathan, Malakhov, and Novikov, 2010, for hedge funds; and Kaplan and Schoar, 2004, for private equity). Nonetheless, the expectation for most types of investors and funds is little persistence in performance.

To analyze persistence we regress various measures of performance (*Revenues*, *Revenues per MSEK Traded*, *Returns*, and the *Sharpe Ratio*) on their lagged values, on both the daily and monthly frequency (except for the *Sharpe Ratio*, which is available on the monthly frequency only). Measures of performance are standardized in the cross section. That is, on each day, firm-level performance is centered on the mean and scaled by the standard deviation across firms. The standardization controls for potential time-variation in the mean and variance of returns. We also estimate persistence regressions with rank-order performance, ranking the relative performance of firms as 1, 2, 3, etc., for each measure of performance on each day or month. We estimate the following regression using OLS:

$$Performance_{i,t} = \beta\, Performance_{i,t-1} + \epsilon_{i,t}, \tag{3}$$

where $Performance_{i,t}$ is defined as above. Results based on daily observations are reported in Table 9, Panel A, and those using monthly observations are reported in Panel B. Since each measure is standardized

[25] Other papers that study competition in the HFT industry include: Boehmer, Li, and Saar (2015), who study competition between HFTs within three 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 (2015), who analyze entry and exit of HFTs and show that increased HFT competition increases market liquidity. Geroski, Gilbert, and Jacquemin (1990) provide an overview of the literature on imperfect competition.

in the cross-section, the constant term in the regression model is mechanically zero. We refer to β as the persistence coefficient, with $\beta = 1$ meaning perfect persistence and $\beta = 0$ meaning no persistence. Standard errors are dually clustered by firm and day in Panel A and by firm and month in Panel B.

INSERT TABLE 9 ABOUT HERE

We find that HFTs have statistically significant daily persistence coefficients of 0.235 for *Revenues* and 0.387 for *Returns*. On the monthly frequency, we find higher persistence coefficients: 0.631 for *Revenues*, 0.763 for the *Sharpe Ratio*, and 0.446 for *Returns*. Performance is more persistent at the monthly level, which is likely due to the higher idiosyncratic risk in daily observations. The rank order analysis shows similar persistence. Consistent with our earlier argument that *Revenues per MSEK Traded* may be a less relevant performance metric for HFTs, we find lower persistence in this measure.

In summary, we find that the performance of HFTs is relatively stable over time, in terms of nominal performance as well as in terms of performance rankings. Firms that have done well in the past typically continue to outperform their competitors in the future.

B. Persistence in Industry-Wide Performance

Having found evidence of persistence in firm-level performance we examine overall performance of the HFT industry over our five year sample. Given that the HFT industry is relatively new, we may observe decreasing industry-wide performance if competition is increasing over time. However, in a market with competition on relative speed, performance may not change. Budish et al. (2015), for example, argue that if HFTs compete on relative latency, regardless of how fast the market as a whole becomes, there will always be a relatively fastest firm that can use its speed advantage to adversely slower traders, capturing a stable level of rents regardless of nominal speed differences. Furthermore, if competition on relative speed makes new entry difficult, as we show in Section VI.D, difficulty of new entry may keep aggregate HFT performance from declining.

Consistent with these predictions, we find that the HFT industry-wide performance is relatively stable over the five-year sample. Table 10, Panel A, reports average daily statistics aggregated across all HFTs and all stocks and reported in half-year intervals. The statistics include *Total Daily Revenues* (*Revenues* summed across all HFTs), *Average Daily Revenues* (the *Total Daily Revenues* averaged across HFTs), *Average Revenues per MSEK Traded* (the ratio of *Total Daily Revenues* and *Daily Trading Volume*), and *Daily Trading Volume* (the *Trading Volume* summed across all HFTs and reported in MSEK). Time trends for *Total Daily Revenues* and *Average Daily Revenues* are also plotted in Figure 2, Panels A and B.

INSERT TABLE 10 ABOUT HERE

INSERT FIGURE 2 ABOUT HERE

Our results show that *Total Daily Revenues* and *Average Daily Revenues* are relatively stable over the five-year period.[26] Interestingly, as shown in Figure 2, Panel B, *Daily Trading Volume per Firm* trends up while *Average Revenues per MSEK Traded* trends down, but the ratio of the two, *HFT Revenues per Firm*, is stable. One possible interpretation is that as HFTs are presumably 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 formally test the movement of the trends, we estimate the following OLS regression on daily observations:

$$Performance_t = \alpha + \beta\,(year - 2010) + \epsilon_t, \tag{4}$$

where $Performance_t$ is one of the performance measures described above, and *year* is a continuous variable (e.g., *year* would take on the approximate value of 2014.25 on March 31, 2014). A positive (negative) coefficient on the $(year - 2010)$ variable corresponds to an increasing (decreasing) trend in the HFT industry-wide performance over the period 2010-2014. Newey-West standard errors with 30 day lags are used. The coefficient estimates are presented in Table 10, Panel B.

The tests statistically confirm the aforementioned trends. However, there is statistical evidence that trading revenues are slightly increasing over the sample period: a yearly increase of 22,682 SEK for *Total Daily Revenues*, relative to a baseline of 166,484 SEK in the first half of 2010. This increase is however not statistically significant in terms of *Average Daily Revenues*, which takes the number of HFTs in the industry into account.

We also examine time trends in the cost of HFT activities for non-HFTs, which is defined as HFT *Total Daily Revenues* divided by *Non-HFT Trading Volume*, calculated on a daily basis. This measure captures the amount of revenue paid from non-HFTs to HFTs per SEK traded. The cost of HFT activities for non-HFTs is relatively small, varying between 0.113 and 0.426 bps, depending on the month, which is about the same order of magnitude as typical exchange fees (see Appendix Section A6). Whereas the exchange fees are direct costs incurred to exchange members and typically passed on to their clients, the

[26] Returns slightly trend down, but, given the relative stability of revenues per firm, this mechanically must mean that average firm capitalization is slightly increasing over the sample.

cost of HFT activities is extracted indirectly through the trading process. Another way to put the cost of HFT activities into perspective is to compare it to the effective spread, which according to Table 1 is between 2 and 6 bps. That is, the cost of crossing the (half) bid-ask spread is more than ten times higher than the cost of HFT activities for non-HFTs. As seen in Figure 2, Panel C, and formally tested in the last column of Table 10, Panel B, there is a small upward time trend in the cost of HFT activities, starting from around 0.13 bps and increasing by about 0.06 on average per year.

Thus, despite concerns that HFTs competing on relative latency might impose costs on non-HFTs, the average observed cost of HFT activities for non-HFTs is surprisingly small.

C. HFT Market Concentration over Time

We next turn to investigate HFT market concentration. While high concentration by itself need not signify lack of competition, there may be reasons why high concentration might be of concern in the context of HFTs. For example, a high concentration among market intermediaries could adversely affect market stability. For example, a dominant HFT firm that incurs a technical malfunction, suffers large capital losses, or suddenly withdraws its liquidity supply could lead to market fragility.

We show that the HFT industry concentration remains high and relatively constant over the five-year sample. We calculate Herfindahl indices, a commonly used measure of concentration of market share or earnings within an industry, for both *Revenues* and *Trading Volume*. The *Herfindahl Industry Concentration by Revenues* is calculated as:

$$Herfindahl_{i,t} = \sum_{i=1}^{N} \left[\frac{Revenues_{i,t}}{HFT\ Revenues_t}\right]^2, \quad (5)$$

where N is the number of firms in month t that earn non-negative trading revenues, $Revenues_{i,t}$ is firm i's total trading revenues in month t, and $HFT\ Revenues_t$ is the trading revenues summed across all HFT firms. *Herfindahl Industry Concentration by Volumes* is calculated using the same formula but considering *Trading Volume* instead of trading *Revenues*. A larger index implies a more concentrated industry: the index is at its minimum $1/N$ when all HFT firms have the same share of the industry and at its maximum of 1 when all activity is concentrated to one firm.

As with the trading performance time trends, the Herfindahl indices are first calculated for each trading day, then averaged across trading days of each half-year block in the data set. Standard deviations are reported in parentheses. The results are reported in Table 10, Panel A, and graphed in Figure 2, Panel D.

The Herfindahl industry concentration in terms of volume lies in the range from 0.186 to 0.304. In terms of revenues, the corresponding interval is from 0.275 to 0.354. As a reference, Van Ness, Van Ness, and Warr (2005) find that market makers have an average Herfindahl index in terms of trading volume on NASDAQ of 0.14, with a range of 0.037 to 0.439.

We test for a time trend in industry concentration by estimating the following OLS regression:

$$Concentration_t = \alpha + \beta\ (year - 2010) + \epsilon_t, \tag{6}$$

where *year* is a continuous variable and $Concentration_t$ is one of the concentration measures discussed above. The coefficient estimates are reported in Table 10, Panel B.

We find no statistically significant time trend for revenues concentration. There is a statistically significant decrease in trading volume concentration, but it is not economically significant. The Herfindahl industry concentration by volume decreases by 0.009 per year from its starting value of 0.292 in the first-half of 2010. The overall conclusion with respect to industry concentration is however that it is relatively constant over the five years.

D. Entry and Exit

To investigate competition in the HFT industry, we examine the properties of firms that just entered the market. New entrants can potentially introduce competition and drive down both firm-level and industry-wide profits. However, frictions in labor markets may prevent the movement of human capital and technical knowledge to new HFTs. For instance, technological expertise and trading strategies are closely guarded trade secrets and employees often agree to non-compete and non-disclosure agreements, preventing them from simply starting new firms or moving to competitors. If new entrants cannot simply pay to acquire skill but require 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 has been active in a given stock, designating less than 1 month, 2 months or 3 months as indicators of new entry.

We estimate the following regression model using OLS:

$$\begin{aligned} Performance_{i,j,t} = {} & \beta_1 \mathbf{1}_{\text{one-month } i,j,t} + \beta_2 \mathbf{1}_{\text{two-month } i,j,t} + \beta_1 \mathbf{1}_{\text{three-month } i,j,t} \\ & + \text{(day x stock)-FEs} + \epsilon_{i,j,t}, \end{aligned} \tag{7}$$

where $Performance_{i,j,t}$ 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 three months

of the sample, since new entry during this period cannot be established. The notation $\mathbf{1}_{one-month\ i,j,t}$ takes the value 1 if firm *i* began trading stock *j* in the last 30 calendar days, otherwise it takes the value 0; similarly, the $\mathbf{1}_{two-month\ i,j,t}$ dummy corresponds to beginning trading in stock *j* in the last 31-60 days, and the $\mathbf{1}_{three-month\ i,j,t}$ dummy corresponds to the previous 61-90 days.[27] If new entrants are less competitive and perform worse than established firms, the coefficient on the one-month dummy should be negative, and, consistent with experience mattering, the two- and three-month dummy coefficients should also be negative though less so. The results are reported in Table 11. Standard errors are dually clustered by firm-stock and by month and are reported within parentheses.

INSERT TABLE 11 ABOUT HERE

Looking at *Revenues*, HFTs have statistically significant negative coefficients corresponding to the new entry dummies, with the negative value of -1,900 SEK for the 1-month dummy, -3,050 SEK for 2 months, and -780 SEK for 3 months. *Returns* are also significantly lower for new entrants in the first and second month. However, *Revenues per MSEK Traded* is statistically insignificant for all three time periods. The fact that the new firms have lower total trading revenues but no statistically significant difference in *Revenues per MSEK Traded* is consistent with HFTs 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 Eq. (7) with the dummy $Exit_{i,j,t}$ as the dependent variable, which takes the value 1 on day *t* for firm *i* if that is the last day firm *i* trades stock j. In this analysis we exclude observations in December 2014, the last month of the analysis, as we are unable to determine market exit in that month. The model is estimated as a linear probability regression, and the results are in Table 11.

We find that, on average, a new entrant in a stock has a higher probability of exit compared to more established HFTs. The effect is found in the one- and two-month indicator variables, both being statistically significant at the 1% level. The daily probability of exit is a statistically significant and economically large amount of 1.4% in the first month. The increased probability of exit is for *each day*, not the cumulative probability of exit at any future point, making the magnitude substantial.

Why do new entrants typically perform worse than established firms? Given our previous findings on latency, we examine whether new entrants are slower than the incumbent HFTs. Repeating the analysis

[27] In accounting for entry and exit, we ignore gaps and just 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 two in our sample) are also not counted as entry/exit events.

in Eq. (7) but replacing the dependent variable with the *Decision Latency* measure (expressed in milliseconds), we find statistically significant coefficients on the 1-month, 2-month, and 3-month dummies of 44.36, 134.6, and 22.8 milliseconds.[28] The positive coefficients on the indicator variables show that new entrants are slower than established firms, which might in part explain their lower trading revenues.

In summary, we find that new entrants tend to perform worse and are more likely to exit than established HFTs. This finding helps explain the continued concentration of revenues among a small subset of established firms: if new entrants tend to exit, then the concentration of revenues continues. A high and steady industry concentration combined with strong firm-level persistence of *Revenues* and *Returns* suggests that top performing incumbent HFTs maintain their position in the market. Together, the four measures of industry structure we examine point to high concentration, consistent with relative latency driving performance.

VII. 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 HFTs. 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. While latency decreases substantially over our five-year sample period, we show that it is relative latency, not nominal latency, that helps explain the differences in performance across HFTs.

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 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 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 non-declining, as is HFT concentration, and new HFT entrants tend to be slower, underperform, and are more likely to exit.

References

[28] 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.

Aït-Sahalia, Y. and M. Saglam, 2014, High-frequency traders: Taking advantage of speed, *Working Paper*.

Biais, B., T. Foucault, and S. Moinas, 2015, Equilibrium fast trading, *Journal of Financial Economics*, *116* (2), 292-313.

Boehmer, E., D. Li, and G. Saar, 2015, Correlated high-frequency trading, *Working Paper*.

Bongaerts, D. and M. Van Achter, 2015, High-frequency trading and market stability, *Working Paper*.

Breckenfelder, J., 2013, Competition between high-frequency traders, and market quality, *Working Paper*.

Brogaard, J. and C. Garriott, 2015, High-frequency trading competition, *Working Paper*.

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan, 2015, Trading fast and slow: Colocation and liquidity, *Review of Financial Studies*, 28 (12), 3407-3443.

Brogaard, J., T. Hendershott, and R. Riordan, 2014, High-frequency trading and price discovery, *Review of Financial Studies*, 27 (8), 2267-2306.

Budish, E., P. Cramton, and J. Shim, 2015, The high-frequency trading arms race: Frequent batch auctions as a market design response, *Quarterly Journal of Economics*, 30 (1547), 1621.

Carhart, M., 1997, On persistence in mutual fund performance, *Journal of Finance*, 52 (1), 57-82.

Cartea, A. and J. Penalva, 2012, Where is the value in high frequency trading?, *Quarterly Journal of Finance*, 2 (3), 1250014.

Chaboud, A., B. Chiquoine, E. Hjalmarsson, and C. Vega, 2014, Rise of the machines: Algorithmic trading in the foreign exchange market, *Journal of Finance*, 69 (5), 2045-2084.

Chen, J., H. Hong, M. Huang, and J. Kubik, 2004, Does fund size erode mutual fund performance? The role of liquidity and organization, *American Economic Review*, 94 (5), 1276-1302.

Clark-Joseph, A., 2013, Exploratory trading, *Working Paper*.

Ellis, K., Michaely, R., & O'Hara, M. (2000). The accuracy of trade classification rules: Evidence from Nasdaq. *Journal of Financial and Quantitative Analysis*, 35(04), 529-551.

Fama, E., and French, K., 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, 33 (1), 3-56.

Fama, E., and J. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy*, 81 (3), 607-636.

Foucault, T., J. Hombert, and I. Roşu, 2016, News trading and speed, *Journal of Finance*, 71 (1) 335-382.

Foucault, T., R. Kozhan, and W. Tham, 2015, Toxic arbitrage, *Working Paper*.

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

Geroski, P., R. Gilbert, and A. Jacquemin, 1990, Barriers to entry and strategic competition, Vol. 41, Taylor & Francis.

Hagströmer, B. and L. Nordén, 2013, The diversity of high-frequency traders, *Journal of Financial Markets*, *16* (4), 741-770.

Hagströmer, B., L. Nordén, and D. Zhang, 2014, How Aggressive Are High-Frequency Traders?, *Financial Review*, *49* (2), 395-419.

Hasbrouck, J., 2003, Intraday price formation in US equity index markets, *Journal of Finance*, 58 (6), 2375-2400.

Hendershott, T., and R. Riordan, 2013, Algorithmic trading and the market for liquidity, *Journal of Financial and Quantitative Analysis*, 48 (4), 1001-1024.

Ho, T. and H. Stoll, 1983, The dynamics of dealer markets under competition, *Journal of Finance*, *38* (4), 1053-1074.

Hoffmann, P., 2014, A dynamic limit order market with fast and slow traders, *Journal of Financial Economics*, *113* (1), 156-169.

Jagannathan, R., A. Malakhov, and D. Novikov, 2010, Do hot hands exist among hedge fund managers? An empirical evaluation, *Journal of Finance*, 65 (1), 217-255.

Jovanovic, B. and A. Menkveld, 2015, Middlemen in limit-order markets, *Working Paper*.

Kaplan, S. and A. Schoar, 2004. Private equity performance: returns, persistence, and capital flows, *Journal of Finance*, 60 (4), 1791-1823.

Kirilenko, A., A. Kyle, M. Samadi, and T. Tuzun, 2015, The flash crash: The impact of high frequency trading on an electronic market, *Working Paper*.

Lee, C. and M. Ready, 1991, Inferring trade direction from intraday data, *Journal of Finance*, *46* (2), 733-746.

Menkveld, A. J. and M. Zoican, 2015, Need for speed? Low latency trading and adverse selection, *Working Paper*.

Petersen, M., 2009, Estimating standard errors in finance panel data sets: Comparing approaches, *Review of Financial Studies*, 22 (1), 435-480.

Roşu, I., 2015, Fast and Slow Informed Trading, *Working Paper*.

Securities and Exchange Commission (U.S.), 2010, Concept release on equity market structure.

Shkilko, A. and K. Sokolov, 2016, Every cloud has a silver lining: Fast trading, microwave connectivity and trading costs, *Working Paper*.

Thompson, S., 2011, Simple formulas for standard errors that cluster by both firm and time, *Journal of Financial Economics*, 99 (1), 1-10.

van Kervel, V. and A. J. Menkveld, 2015, High-frequency trading around large institutional orders, *Working Paper*.

Van Ness, B., R. Van Ness, and R. Warr, 2005, The impact of market-maker concentration on adverse selection costs for NASDAQ stocks, *Journal of Financial Research*, 28 (3), 461-485.

Weller, B., 2013, Intermediation chains and specialization by speed: Evidence from commodity futures markets, *Working Paper*.

Weston, J., 2000, Competition on the Nasdaq and the impact of recent market reforms, *Journal of Finance*, *55* (6), 2565-2598.

Yao, C. and M. Ye, 2015, Why trading speed matters: a tale of queue rationing under price controls, *Working Paper*.

Yueshen, B. Z., 2014, Queuing uncertainty in limit order market, *Working Paper*.

Figure 1: HFT *Decision Latency* over time

This figure plots *Decision Latency* (indicated on the vertical axis on a log scale) from January 2010 to December 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 one second. *HFT #1* is the *Decision Latency* of the fastest HFT firm in each month; *HFTs #1-5* is the average *Decision Latency* of the 5 fastest HFTs in each month; and *All HFTs* is the average *Decision Latency* across all HFTs for which latency is lower than one 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.

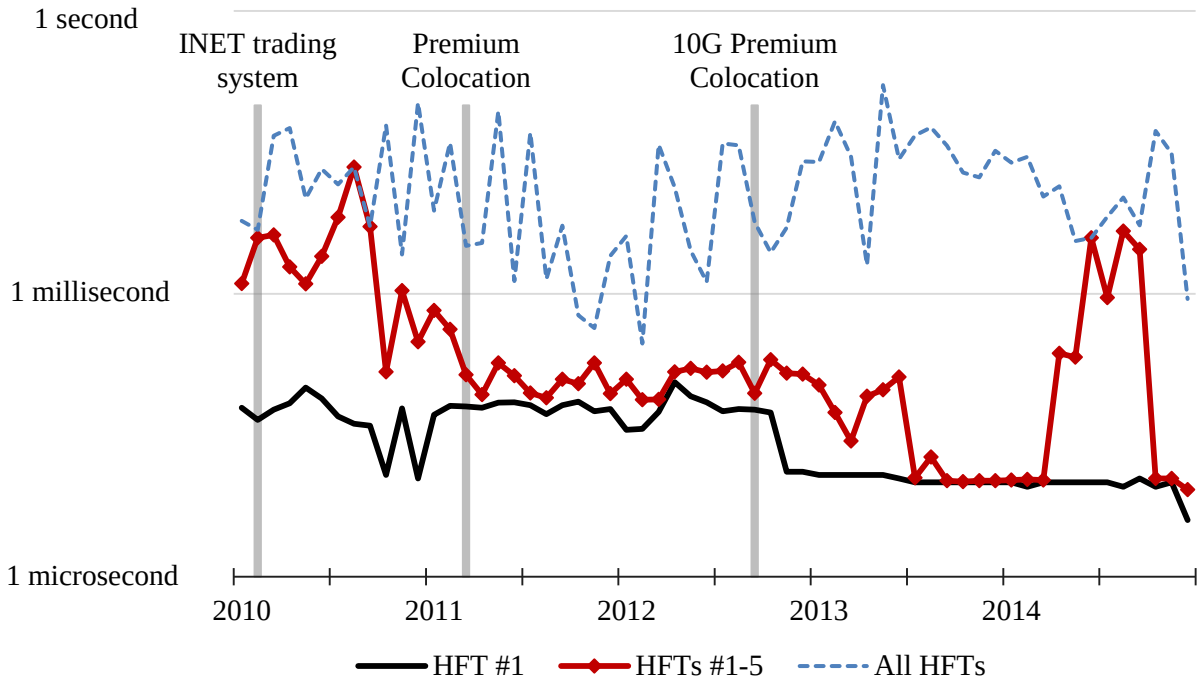
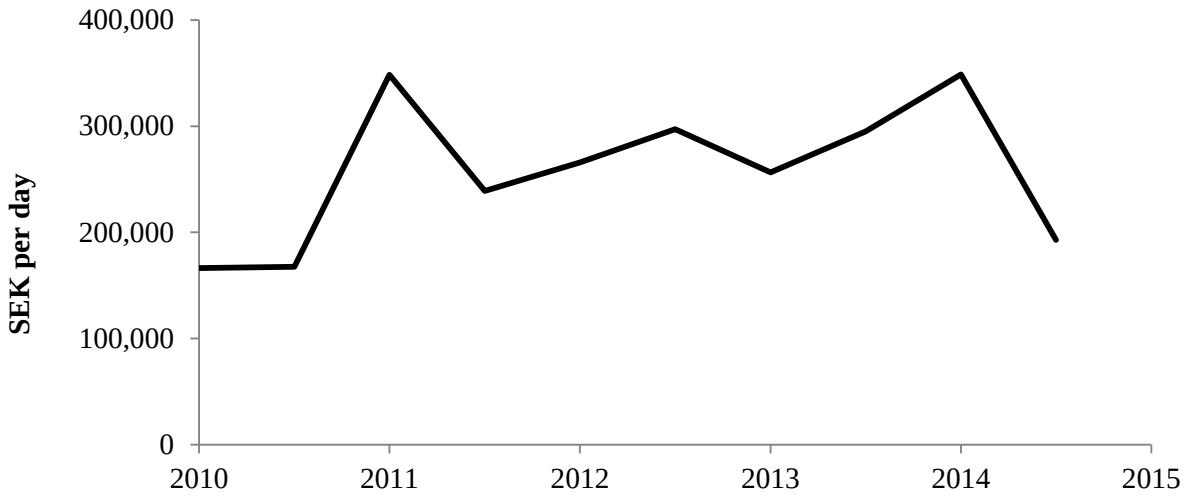


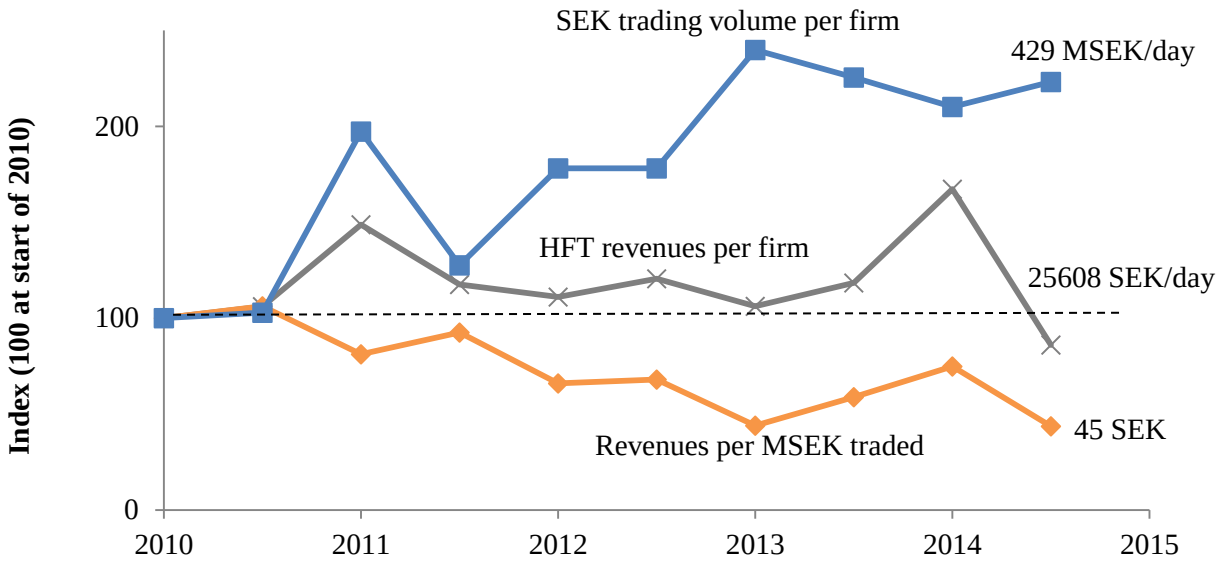
Figure 2: HFT revenues, cost of HFT activities, and industry concentration over time

This figure shows various time series trends. Panel A plots *Total Daily Revenues*, which is the average daily sum of *Revenues* (defined as in Table 2) across all HFT firms. Panel B 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*, which are also plotted. Panel C plots the *Cost of HFT Activities to Non-HFTs*, calculated as HFT *Total Daily Revenues* divided by the total non-HFT trading volume (in SEK). Panel D plots two indexes of *HFT Industry Concentration*: Herfindahl indexes calculated on daily *Revenues* or with daily *Trading Volumes*, see Eq. (5). The sample consists of 25 Swedish stocks and 60 months of trading (January 2010 to December 2014). The time series are reported on a biannual frequency.

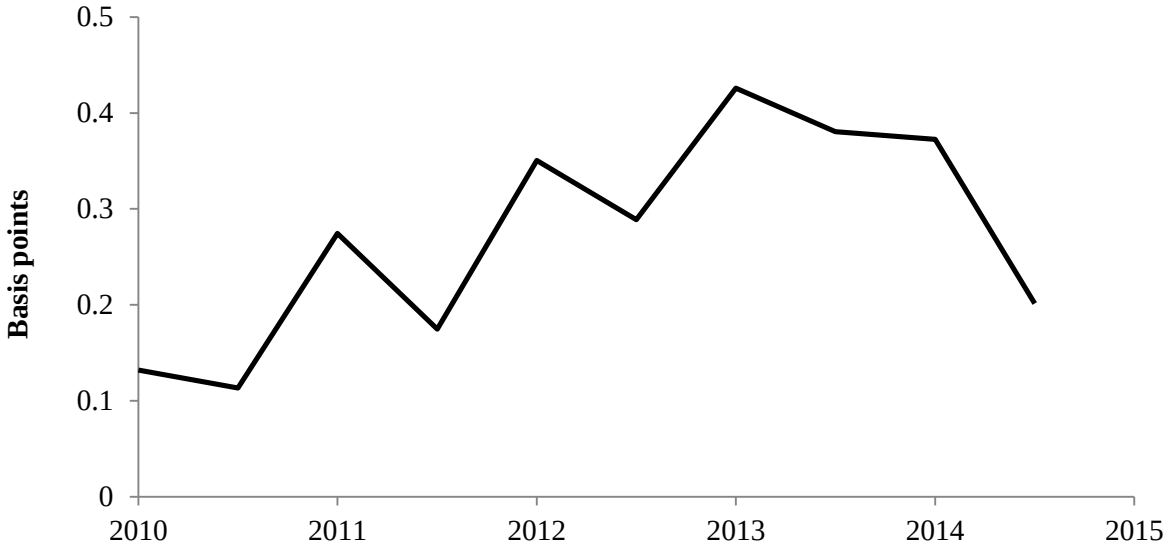
Panel A: Total Daily Revenues



Panel B: Firm-Average Daily Revenues



Panel C: Cost of HFT Activities to Non-HFTs



Panel D: HFT Industry Concentration

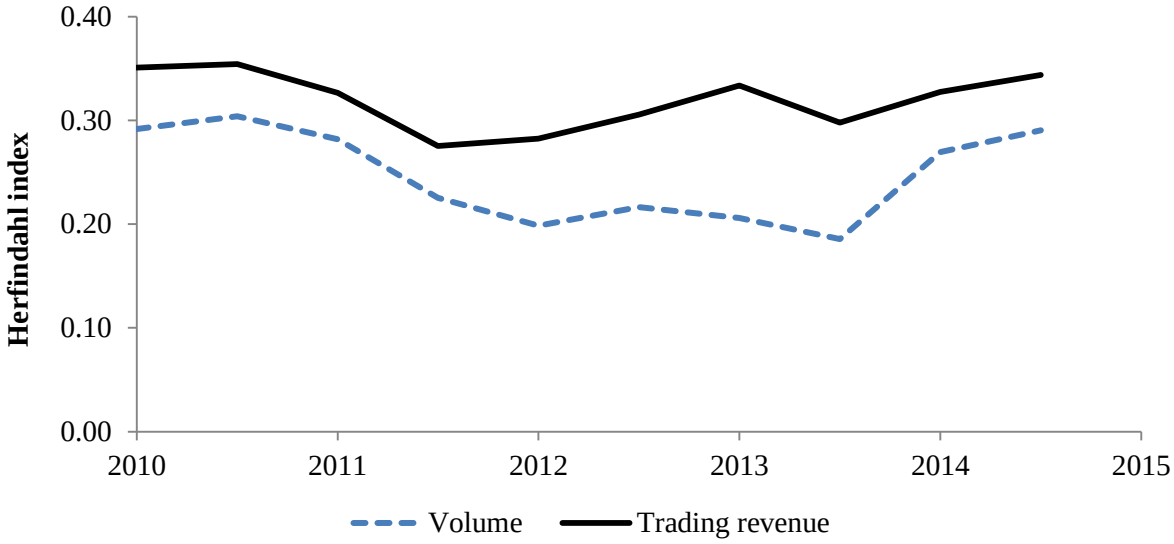


Table 1: Stock characteristics

This table reports summary statistics of the 25 Swedish stocks in the sample. *Market Capitalization* is based on the closing price on December 31, 2014 (expressed in MSEK). All other statistics are calculated as averages across trading days in December 2014. Because *SCVb* is delisted in May 2014, the metrics for that stock are based on April 2014. *Daily Trading Volume* refers to trading at NASDAQ OMX Stockholm only and is reported in MSEK. *Daily Turnover* is the Daily Trading Volume divided by Market Capitalization, expressed in percentage points. *Tick Size* is the average minimum price change; *Quoted Spread* is the average bid-ask spread prevailing just before each trade; and *Effective Spread* is the trade value-weighted average absolute difference between the trade price and the bid-ask midpoint. All spread measures are based on continuous trading at NASDAQ OMX Stockholm, expressed relative to the bid-ask spread, and presented in basis points. The Tick Size and the Quoted Spread are halved to be comparable to the Effective Spread. *Volatility* is the average 10-second squared returns, calculated from bid-ask midpoints. The *Fragmentation Index* is the inverse of a Herfindahl index of trading volumes across the five largest trading venues (BATS, Burgundy, Chi-X, NASDAQ OMX Stockholm, and Turquoise); a higher value signifies greater fragmentation. The table is sorted by Market Capitalization.

Stock Ticker	Market Cap. (MSEK)	Daily Trading Vol. (MSEK)	Daily Turnover (%)	Tick Size (bps)	Quoted Spread (bps)	Effective Spread (bps)	Volatility (sq. bps)	Fragment-ation Index
HMb	475,595	1,358	0.29	1.57	2.06	2.26	2.94	2.11
SHBa	228,731	757	0.33	1.38	2.17	2.36	4.08	2.30
SWEDa	221,307	1,011	0.46	2.58	2.90	3.21	5.69	2.00
SEBa	216,025	771	0.36	2.67	3.19	3.30	5.20	1.93
ATCOa	183,324	839	0.46	2.32	3.04	3.23	5.12	2.29
ASSAb	145,878	529	0.36	1.23	2.36	2.51	3.52	2.28
VOLVb	136,816	1,099	0.80	2.98	3.24	3.34	5.71	1.91
INVEb	129,676	590	0.46	1.79	2.43	2.62	4.12	1.80
SCAb	104,559	632	0.60	2.91	3.42	3.59	5.13	2.13
SAND	95,835	798	0.83	3.27	3.67	3.74	6.55	2.12
SCVb	78,800	825	1.05	2.65	3.66	4.75	6.75	1.91
ATCOb	78,395	262	0.33	2.55	3.88	4.09	5.52	1.85
SKFb	68,879	684	0.99	3.14	3.63	3.85	6.07	2.24
ELUXb	68,807	471	0.68	2.24	3.16	3.29	5.05	2.31
SKAb	67,160	320	0.48	3.07	3.63	3.77	4.89	1.90
ALFA	62,205	463	0.75	3.42	4.06	4.10	6.04	2.05
KINVb	60,097	337	0.56	1.95	3.61	3.98	7.95	2.26
SWMA	49,082	337	0.69	2.04	3.10	3.32	4.38	2.24
TEL2b	40,403	334	0.83	2.64	3.21	3.51	5.97	2.12
GETIb	39,540	221	0.56	2.91	3.81	4.21	4.25	1.99
LUPE	34,964	586	1.68	4.08	5.65	5.68	16.53	1.91
BOL	34,326	538	1.57	4.05	4.87	4.90	7.88	2.06
SECUb	32,861	169	0.51	2.73	3.66	3.88	3.75	2.32
MTGb	15,369	155	1.01	2.05	4.07	4.77	7.46	2.17
SSABa	13,877	370	2.67	1.61	3.82	4.14	8.89	1.76

Table 2: The cross-section of HFT performance

This table reports descriptive statistics on 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; 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 the 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 one 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 (January 2010 to December 2014).

	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
Aggressive 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

(N = 16 firms)

Table 3: Trading revenues, costs, and profits from public filings

This table reports *Trading Revenues*, *Trading Costs*, *Trading Profit Margins*, and *Trading Returns* calculated from annual reports and IPO prospectuses for five high-frequency trading firms for which public data is available. *Trading Costs* are broken down into several categories, all expressed as a percent of trading revenues. *Trading Profit Margin* is calculated as (1 - *Trading Costs*), and *Trading Returns* are calculated two ways based on two capitalization measures: as *Trading Revenues / (Trading Assets Minus Trading Liabilities)* and as *(Trading Revenues / Book Equity)*. All quantities are in million USD, except for the firm Flow Traders, which is in million EUR.

	Virtu					KCG			GETCO				Flow Traders				Jump
	2015	2014	2013	2012	2011	2015	2014	2013	2012*	2011	2010	2009	2015	2014	2013	2012	2010
Trading Revenues (in millions)	757.5	685.2	623.7	581.5	449.4	1,179.9	1,274.4	903.8	526.6	896.5	865.1	955.2	400.1	240.8	200.5	125.1	511.6
-- % of revenue from proprietary trading	95.1%	98.5%	98.4%	100%	100%	73.8%	68.5%	67.0%	89.9%	94.2%				100%	100%	100%	
Trading Costs (as % of Trading Revenue)	54.7%	60.0%	57.8%	72.6%	62.1%	51.3%	52.4%	59.0%	62.5%	48.5%	48.6%	40.4%	35.5%	41.6%	43.7%	47.5%	
-- Brokerage, exchange and clearance fees	30.7%	33.7%	31.3%	34.5%	32.9%	22.5%	23.9%	27.3%	35.3%	32.2%	35.1%	32.1%	{ 14.2%	15.7%	15.8%	14.8%	
-- Communication and data processing	9.0%	10.0%	10.4%	9.5%	10.3%	11.8%	11.8%	13.7%	17.2%	9.7%	7.1%	4.5%					
-- Equipment rentals, depreciation and amortization	4.4%	4.5%	4.0%	15.7%	11.1%	10.2%	10.4%	11.0%	9.1%	6.2%	6.2%	3.8%	1.6%	1.8%	1.9%	2.4%	
-- Net interest & dividends on securities paid (on credit lines, securities borrowing, etc.)	7.1%	8.6%	7.8%	7.1%	6.0%	6.1%	5.4%	6.5%	1.0%	0.3%	0.1%	0.0%	9.7%	12.5%	12.8%	12.3%	
-- Other trading costs (administrative & technical costs, other overhead, etc.)	3.4%	3.2%	4.4%	5.8%	1.8%	0.7%	0.8%	0.5%	0.0%	0.0%	0.0%	0.0%	10.1%	11.5%	13.3%	18.0%	
Trading Profit Margin	45.3%	40.0%	42.2%	27.4%	37.9%	48.7%	47.6%	41.0%	37.5%	51.5%	51.4%	59.6%	64.5%	58.4%	56.3%	52.5%	52.3%***
Trading Revenue / (Trading Assets Minus Trading Liabilities)**	183%	228%	196%	184%		96%	96%	60%	62%				114%	118%	119%	103%	237%
Trading Revenue / (Book Equity)	136%	135%	138%	84%		82%	84%	60%	80%				162%	169%	146%	123%	222%

* Does not include any costs associated with the December 19, 2012 merger agreement with Knight, such as any costs related to the Knight August 1, 2012 incident

** Trading asset include cash and cash equivalents, financial instruments owned, receivables from broker-dealers and clearing organizations, and collateralized agreements. Trading liabilities include short-term borrowings, collateralized financing, financial instruments sold and not yet purchased, payables to broker-dealers and clearing organizations, and other accounts payable.

*** Total profit margin

Table 4: Trading performance and latency

This table analyzes the relationship between trading performance and latency. It reports coefficients estimated from Eq. (1) for five performance measures as dependent variables: *Revenues*, *Returns*, *Sharpe Ratio*, *Trading Volume*, and *Revenues per MSEK Traded* (all defined as in Table 2). *Revenue*, *Returns*, *Revenues per MSEK Traded*, and *Trading Volume* are calculated at a daily level for each HFT firm aggregated across stocks and then averaged across trading days in each month to get firm-month observations; and the *Sharpe Ratio* is calculated as firm-month observations using the mean and standard deviation of daily observations of *Revenues* aggregated across stocks. We estimate OLS regressions with month fixed effects. The independent variables considered are as follows: log(*Decision Latency*) is the natural logarithm of *Decision Latency* (defined as in Table 2). *Top 1* and *Top 1-5* 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. The *End-of-Day Inventory Ratio*, *Max. Intraday Inventory Ratio*, *Investment Horizon*, and the *Aggressiveness Ratio* are defined as in Table 2. 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 *** correspond to statistical significance at the 10%, 5%, and 1%, respectively. Standard errors are dually clustered by firm and month and are reported in parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (January 2010 to December 2014).

	Revenues			Returns			Sharpe Ratio			Trading Volume (x 10^{-6})			Revenues per MSEK Traded		
log(Decision Latency)	-14020***	-1063	9925	-.221***	-.059	-.00349	-4.38***	-1	2.03	-247***	-89.7	10.5	-19.4	-10.7	101**
	(4311)	(6358)	(10481)	(.0483)	(.065)	(.0852)	(.632)	(1.2)	(1.46)	(43.7)	(59.1)	(74)	(57.5)	(69.1)	(40.4)
Top 1		29849*	24639**		.238*	.252*		3.77*	4.2*		326***	281***		6.99	57.6*
		(15251)	(12249)		(.134)	(.142)		(2.21)	(2.29)		(97.9)	(104)		(51.7)	(32.8)
Top 1-5		24074**	15451*		.333**	.303**		7.29**	5.61**		301**	201**		19.4	44.1
		(11619)	(8009)		(.155)	(.133)		(3.24)	(2.63)		(132)	(97.4)		(93.3)	(55.9)
End-of-Day Inv.			2921			.0839*			2***			-33.9**			326*
			(3774)			(.0494)			(.74)			(15.9)			(168)
Max Intraday Inv.			-21008**			*[omitted]*			-3.74***			-183***			-76.3
			(8579)						(1.23)			(65.3)			(127)
Investment Horizon			-5401			-.134***			-2.25***			-76.4			-73.3
			(5994)			(.0404)			(.726)			(50.3)			(63.8)
Aggressive Ratio			5481			-.0212			-.779			41.7			-55.5
			(3865)			(.0558)			(.823)			(28.8)			(65.8)
Constant	20278***	8466**	10894**	.254***	.104*	.107**	5.1***	1.94	2.26*	313***	169***	198***	35.2	27	7.91
	(6973)	(4189)	(4885)	(.0579)	(.0587)	(.0513)	(1.26)	(1.23)	(1.23)	(75.9)	(57.8)	(56.3)	(57.3)	(80.2)	(10)
Month FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	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

Table 5: Relative latency and trading performance around colocation upgrades

This table examines the relationship between trading performance and latency around two colocation upgrades on March 14, 2011 and September 17, 2012. Specifically, it analyzes two groups of 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 *Revenue, Returns, Revenues per MSEK Traded,* and *Trading Volume* are calculated at a daily level for each HFT and then averaged across trading days and HFTs 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 six degrees of freedom).

	Revenues				Returns				Sharpe Ratio			
HFT latency rank	Before	After	Diff.	(S.E.)	Before	After	Diff.	(S.E.)	Before	After	Diff.	(S.E.)
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)
Diff-in-diff			**41,978*** **	(6,533)			**0.165** **	(0.045)			**1.01***	(0.50)

	Trading Volume (x 10^{-6})				Revenues per MSEK Traded			
HFT latency rank	Before	After	Diff.	(S.E.)	Before	After	Diff.	(S.E.)
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)
Diff-in-diff			**171.9** **	(50.1)			**34.1**	(40)

Table 6: Alternative latency measures

This table is similar to Table 4 but analyzes two alternative latency measures: *Queuing Latency* (Panel A), which counts the cases where an HFT firm captures the top-of-queue position in a limit-order book gap; and *Mean Latency* (Panel B), the mean of a distribution of latencies in each firm-month where a passive trade is followed by an active trade at the same venue, in the same stock, within one millisecond. The construction of these alternative latency measures and their relationship to HFT strategies is discussed in Section IV.C. As in Table 4, the table reports coefficients estimated from Eq. (1) for five performance measures as dependent variables: *Revenues*, *Returns*, *Sharpe Ratio*, *Trading Volume*, and *Revenues per MSEK Traded*. We estimate OLS regressions with month fixed effects. In Panel A, log(*Queuing Latency* + 1) is the natural logarithm of *Queuing Latency* + 1, and, in Panel B, log(*Mean Latency*), is the natural logarithm of *Mean Latency*. *Top 1* and *Top 1-5* are indicator variables for whether a given firm is ranked among the top 1 or top 1-5 firms by the corresponding latency measure. All other independent variables are as in Table 4. *, ** and *** correspond to statistical significance at the 10%, 5%, and 1%, respectively. Standard errors are dually clustered by firm and month and reported in parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (January 2010 to December 2014).

Panel A: *Queuing Latency* and trading performance

	Revenues			Returns			Sharpe Ratio			Trading Volume (x 10^{-6})			Revenues per MSEK Traded		
log(Queuing Latency + 1)	16761***	4150	-8265	.281***	.176***	.121*	5.43***	2.94***	-.283	288***	133**	34.8	32.9	38.1	-95
	(4608)	(4907)	(11994)	(.0533)	(.0529)	(.0671)	(.72)	(1.12)	(1.43)	(39.6)	(56.7)	(102)	(58.6)	(77.1)	(62.1)
Top 1		50803***	51684***		.463**	.471**		11.9***	12.6***		601***	603***		-9.21	35.5
		(18676)	(15865)		(.228)	(.218)		(2.4)	(2.13)		(126)	(119)		(75.9)	(75.6)
Top 1-5		13698*	9563		.102	.127		2.13	2.39		176*	128		-9.11	87.7
		(7180)	(7400)		(.123)	(.136)		(1.74)	(1.88)		(99.9)	(87.4)		(125)	(66.2)
End-of-Day Inv.			2718			.0858*			1.95**			-32.5*			325*
			(3742)			(.0508)			(.767)			(18.1)			(170)
Max Intraday Inv.			-19571**						-2.9**			-151**			-58.6
			(8658)						(1.25)			(68.5)			(129)
Investment Horizon			-4950			-.0917***			-1.96**			-60.5			-72.8
			(7114)			(.0333)			(.846)			(60.8)			(62.1)
Aggressive Ratio			6664			.00695			-.442			60.3*			-51.4
			(4551)			(.0428)			(.617)			(33.9)			(65.8)
Constant	20223***	10893**	11153**	.252***	.176***	.16***	5.08***	3.32***	2.91**	313***	197***	203***	34.8	39.1	-7.21
	(6685)	(5372)	(4814)	(.0509)	(.0557)	(.0496)	(1.13)	(1.28)	(1.19)	(72.2)	(67.7)	(58.2)	(56.6)	(96.2)	(45.5)
Month FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.152	0.213	0.302	0.267	0.300	0.323	0.285	0.357	0.429	0.377	0.474	0.547	0.080	0.080	0.146
N	737	737	737	737	737	737	737	737	737	737	737	737	737	737	737

Panel B: *Mean Latency* and trading performance

	Revenues			Returns			Sharpe Ratio			Trading Volume (x 10^{-6})			Revenues per MSEK Traded		
log(Mean Latency)	-10796***	3261	7350	-.14***	.00246	.0247	-3***	-.0674	1.06	-160***	-32.7	17.7	-5.42	22.9	50.3
	(3456)	(6724)	(5858)	(.0523)	(.0918)	(.0913)	(.904)	(1.74)	(1.61)	(58.9)	(93.6)	(77.5)	(42.5)	(54)	(73.4)
Top 1		12076	19574		-.0658	.0144		-2.99	-.038		-220	-116		60.9	151
		(26687)	(20144)		(.274)	(.263)		(4.63)	(4.02)		(216)	(148)		(93.1)	(142)
Top 1-5		33497***	13819		.433***	.301**		9.69***	4.69*		465***	216**		49.4	-40.7
		(10315)	(10857)		(.148)	(.142)		(3)	(2.45)		(115)	(109)		(122)	(71.7)
End-of-Day Inv.			2352			.0737*			1.83**			-40.3**			322*
			(3909)			(.0442)			(.746)			(17.5)			(167)
Max Intraday Inv.			-19668***						-3.65***			-217***			-52.1
			(6100)						(.999)			(51.1)			(117)
Investment Horizon			-4960			-.171***			-2.19***			-82.3*			-69
			(5628)			(.0469)			(.743)			(46.1)			(61.1)
Aggressive Ratio			6898**			-.0214			-.468			54.6**			-48.5
			(3469)			(.0521)			(.764)			(25.3)			(61.5)
Constant	20554***	7620	12494	.258***	.109	.137**	5.19***	1.96	3.14**	319***	170*	230***	35.7	13.3	32.6
	(7076)	(8590)	(8466)	(.0704)	(.0809)	(.0658)	(1.44)	(1.7)	(1.4)	(83.9)	(95.4)	(88.6)	(57.5)	(98.7)	(45)
Month FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.093	0.142	0.253	0.128	0.182	0.236	0.124	0.219	0.345	0.146	0.266	0.423	0.079	0.080	0.147
N	737	737	737	737	737	737	737	737	737	737	737	737	737	737	737

Table 7: Price impact, realized spread and latency

This table examines the relationship between *Decision Latency* and two dependent variables: active *Price Impact* and passive *Realized Spread*. *Price Impact* is the basis point change in spread midpoint from just before to ten seconds after a trade initiated by an HFT firm. *Realized Spread* is the basis point difference between the transaction price and the bid-ask spread midpoint ten seconds after a trade where an HFT firm was the liquidity provider. *Top 1* and *Top 1-5* are indicator variables for whether a given firm is ranked among the top 1 or top 1-5 firms by speed in a given month. *Decision Latency*, *Average End-of-Day Inventory*, *Maximum Intraday Inventory*, *Investment Horizon*, and the *Aggressive Ratio* are defined as in Table 2. We also control for the following stock-month-specific variables: *Volatility*, *Fragmentation Index*, *Tick Size*, *Quoted Spread* (all defined as in Table 1), and *Non-HFT Trading Vol.*, which is the SEK trading volume recorded by non-HFT brokers. All continuous variables are in units of standard deviations. *, ** and *** correspond to p-values lower than 10%, 5%, and 1%, respectively. Standard errors are dually clustered by firm-stock and month and are reported in the parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (January 2010 to December 2014); stock-firm-month observations for which an HFT firm does not trade actively or passively are excluded.

	Price Impact		Realized Spread	
log(Decision Latency)	-.318	-.494*	-.364***	-.384***
	(.225)	(.212)	(.0982)	(.0958)
Top 1	.371*	.337*	.0214	.0599
	(.201)	(.193)	(.131)	(.126)
Top 1-5	.73**	.645**	.448***	.477***
	(.362)	(.315)	(.136)	(.118)
Avg. End-of-Day Inv.		.106		.00393
		(.158)		(.0655)
Max. Intraday Inv.		.256		.0866
		(.216)		(.0663)
Investment Horizon		-.158		-.0467
		(.182)		(.0678)
Aggressiveness Ratio		.0103		-.0544
		(.121)		(.0581)
Non-HFT Trading Vol.		-.100		.145**
		(.162)		(.0618)
Volatility		-.226		-.462**
		(.161)		(.217)
Fragmentation Index		-.162*		-.0943*
		(.0966)		(.0538)
Tick Size		-.0469		-.265*
		(.301)		(.136)
Quoted Spread		.890**		.623***
		(.379)		(.163)
Constant	3.91***	3.96***	-.0958	-.108
	(.182)	(.22)	(.084)	(.107)
(Month x Stock) FEs	Yes	No	Yes	No
R-squared	0.196	0.016	0.158	0.017
N	11449	11449	11269	11269

Table 8: Cross-market arbitrage and latency

This table 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 HFT* firm performs the trade and 0 if a *Slow HFT* does it. In each month, *Slow HFTs* are those who are not among the top five HFTs in terms of trading speed. *Fast HFTs* are either the top 1 or top 1-5 HFTs in terms of trading speed. Trades included in the analysis are those performed by HFTs in the sample stocks. A news event is defined as when the absolute return on the OMXS30 futures during a one-second window preceding the stock trade is "large" (in the top decile for each month among non-zero 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 zero if there is no news event in the one-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 ten seconds; *Lagged Volume*, the *SEK Trading Volume* (divided by 100,000) over the previous ten seconds; *Quoted Spread*, the difference between the best bid and offer quotes (multiplied by 10,000), divided by the midpoint quote; *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 January 2010 to December 2014; reported coefficients and marginal effects are means across the sixty month, and standard errors are in parentheses. *, ** and *** correspond to p-values lower than 10%, 5%, and 1%, respectively.

Panel A: Active Trading

	"Fast" = Top 1 HFT		"Fast" = Top 1-5 HFT	
	Probit (1=Fast HFT)	Marginal effects	Probit (1=Fast HFT)	Marginal effects
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 FEs	Yes		Yes	
Average N	109684		277044	
Avg. psuedo-R^2	0.209		0.169	

Panel B: Passive Trading

	"Fast" = Top 1 HFT		"Fast" = Top 1-5 HFT	
	Probit (1=Fast HFT)	Marginal effects	Probit (1=Fast HFT)	Marginal effects
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 FEs	Yes		Yes	
Average N	95268		258409	
Avg. psuedo-R^2	0.204		0.163	

Table 9: Persistence of HFT performance

This table analyzes persistence in HFTs' performance on both a daily (Panel A) and a monthly (Panel B) frequency. The persistence coefficient reported in the table is the β estimated in the regression model given in Eq. (3), where *Performance* can be one of the following dependent variables calculated on a daily basis for each HFT firm: *Revenues*, *Returns*, *Sharpe Ratio*, and *Revenues per MSEK Traded* (all defined as in Table 2). For the monthly frequency, each of the performance measures defined on a daily frequency are first averaged across trading days within each month. The *Sharpe Ratio* is considered only for the monthly regressions and is based on mean and standard deviation of daily observations of *Revenues*. The variables are either in units of standard deviations for each day or month (in each time period, firm-level performance is centered on the mean and scaled by standard deviation across HFTs) or on the rank order of the HFTs (from 1 to 16 based on performance). *, ** and *** correspond to p-values lower than 10%, 5%, and 1%, respectively. Standard errors are dually clustered by firm and day (or month for Panel B) and are reported in parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (January 2010 to December 2014).

Panel A: Daily persistence

	Standardized			Rank order		
	Revenues	Returns	Revenues per MSEK Traded	Revenues	Returns	Revenues per MSEK Traded
Lag dependent variable	.235***	.387***	.023	.234***	.283***	.029*
	(0.087)	(0.085)	(0.020)	(0.067)	(0.064)	(0.018)
R-squared	0.057	0.157	0.016	0.114	0.143	0.076
N	10642	10642	10642	10642	10642	10642

Panel B: Monthly persistence

	Standardized				Rank order			
	Revenues	Returns	Sharpe Ratio	Revenues per MSEK Traded	Revenues	Returns	Sharpe Ratio	Revenues per MSEK Traded
Lag dependent variable	.631***	.446***	.763***	.106	.464***	.539***	.196**	.134**
	(0.113)	(0.155)	(0.062)	(0.083)	(0.091)	(0.094)	(0.095)	(0.063)
R-squared	0.401	0.222	0.584	0.060	0.252	0.325	0.091	0.069
N	737	737	737	737	737	737	737	737

Table 10: Long-run trends in HFT concentration and industry-wide performance

This table reports long-run trends in various variables related to the HFT industry. *Total Daily Revenues* (*Average Daily Revenues*) is *Revenues* summed (averaged) across all HFTs, reported in SEK; *Average Revenues per MSEK Traded* is daily *Revenues per MSEK Traded* aggregated across HFTs; *Daily Trading Volume* is the daily *Trading Volume* summed across HFTs; *Cost of HFT Activities for Non-HFTs* is HFT *Revenues* divided by non-HFT trading volume. The table also reports long-run trends in various variables measuring the HFT industry concentration in terms of revenues and volumes: specifically, the *Herfindahl Concentration by Revenues* or *by Trading Volumes* is calculated as in Eq. (5). All measures are calculated on a daily frequency and reported in Panel A as the average across the trading days of each half-year period. Standard errors are given in parentheses. Panel A also reports *T*, the number of trading days in each half-year period. Panel B reports estimates of regressions aimed at identifying time trends, specified for each variable observed at daily frequency. The regression specification is given in Eq. (4). An estimated $\beta > 0$ indicates an increasing trend in the dependent variable. *, ** and *** correspond to p-values lower than 10%, 5%, and 1%, respectively. In Panel B, Newey-West standard errors (using 30 day lags) are in parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (January 2010 to December 2014).

Panel A: Biannual averages

	Total Daily Revenues (SEK)	Average Daily Revenues (SEK)	Average Revenues per MSEK Traded (SEK)	Daily Trading Volume (MSEK)	Cost of HFT Activities for Non-HFTs (bps)	Herfindahl Industry Conc. by Trading Volumes	Herfindahl Industry Conc. by Revenues	T (days)
2010:1	166,484	19,694	99.71	1,626	0.132	0.292	0.351	124
	(19439)	(2256)	(10.74)	(57.03)	(0.014)	(0.059)	(0.083)	
2010:2	167,582	20,877	105.81	1,589	0.113	0.304	0.354	130
	(26100)	(3257)	(18.13)	(65.69)	(0.015)	(0.04)	(0.057)	
2011:1	348,282	29,299	80.93	4,512	0.274	0.282	0.327	124
	(79005)	(6577)	(13.36)	(148.62)	(0.049)	(0.025)	(0.067)	
2011:2	239,140	23,156	92.26	2,536	0.175	0.225	0.275	130
	(28225)	(2668)	(9.43)	(80.15)	(0.018)	(0.032)	(0.111)	
2012:1	265,916	21,868	65.78	4,167	0.351	0.199	0.282	123
	(34121)	(2879)	(7.4)	(88.5)	(0.056)	(0.044)	(0.098)	
2012:2	297,164	23,743	67.83	4,287	0.289	0.216	0.306	127
	(44538)	(3490)	(9.86)	(108.34)	(0.038)	(0.026)	(0.109)	
2013:1	256,618	20,898	43.92	5,665	0.426	0.206	0.334	122
	(30104)	(2375)	(4.57)	(108.73)	(0.045)	(0.02)	(0.135)	
2013:2	294,979	23,313	58.70	5,487	0.381	0.186	0.298	128
	(35315)	(2756)	(5.75)	(129.71)	(0.038)	(0.03)	(0.105)	
2014:1	348,687	32,937	74.56	4,280	0.372	0.269	0.328	121
	(69728)	(6069)	(10.27)	(144.08)	(0.059)	(0.024)	(0.105)	
2014:2	192,896	16,945	43.59	4,888	0.201	0.290	0.344	129
	(35573)	(3244)	(7.93)	(153.5)	(0.036)	(0.016)	(0.073)	

Panel B: Time trend regressions

	Total Daily Revenues (SEK)	Average Daily Revenues (SEK)	Average Revenues per MSEK Traded (SEK)	Daily Trading Volume (MSEK)	Cost of HFT Intermedia-tion for Non-HFTs (bps)	Herfindahl Industry Conc. by Trading Volumes	Herfindahl Industry Conc. by Revenues
(year - 2010)	22682*	852.6	-11.61***	788.9***	.05635***	-.0092*	-.0022
	(12130)	(1128)	(3.124)	(117.4)	(.01138)	(.0049)	(.0034)
Constant	199748***	21123***	103.4***	1867***	.1266***	.271***	.325***
	(32304)	(3098)	(10.06)	(302.2)	(.02547)	(.0119)	(.0107)
R-squared	0.004	0.001	0.020	0.361	0.031	0.068	0.001
N	1255	1255	1255	1255	1255	1255	1255

Table 11: HFT entry and exit analysis

This table analyzes the determinants of HFT entry and exit into stocks. The performance measures *Revenues*, *Revenues per MSEK Traded*, and *Returns* are defined as in Table 2 for each HFT firm, stock, and trading day. The table reports coefficient estimates from Eq. (7), estimated on a panel of firm-stock-day observations. The *one-month dummy* takes the value 1 if firm *i* began trading stock *j* in the last 30 days, otherwise 0; *two-* and *three-month dummy* variables are defined similarly. We exclude the observations in January 2010 as this is when we first observe any firm. In the fourth column, a linear probability regression is estimated using Eq. (7) but with the dummy $Exit_{i,t}$ as the dependent variable, which takes the value 1 on day *t* for firm *i* if that is the last day firm *i* trades. This regression excludes observations in December 2014, the last month of the analysis, since we cannot determine exits. In the fifth column, an OLS is estimated on a firm-stock-month panel using Eq. (7) but with *Decision Latency* as the dependent variable. (Since *Decision Latency* is a firm-month variable, it is assigned to be constant across stocks). *, ** and *** correspond to p-values lower than 10%, 5%, and 1%, respectively. Standard errors are dually clustered by firm-stock and month and are reported in the parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (January 2010 to December 2014).

	Revenues (thous. SEK)	Revenues per MSEK Traded	Returns	Daily Probability of Exit (x 10^3)	Decision Latency (in milliseconds, monthly obs.)
One-month dummy	-1.90**	-97.46	-.032**	1.455***	44.36*
	(.93)	(209.7)	(.014)	(.420)	(26.73)
Two-month dummy	-3.05**	-87.11	-.033***	1.486***	134.6***
	(1.434)	(230.3)	(.011)	(.425)	(33.98)
Three-month dummy	-.78	104.7	-.018*	-.194	22.8**
	(1.35)	(196.6)	(.010)	(.477)	(11.19)
Constant	1.43***	76.64***	.017***	.530***	14.87***
	(.19)	(4.12)	(.002)	(.049)	(.89)
Day x Stock FEs	Yes	Yes	Yes	Yes	(Month x Stock FEs)
R-squared	0.101	0.129	0.147	0.154	0.432
N	241053	241053	241053	241053	11014

Appendix

A1. TRS data processing

This appendix presents a summary of our data processing. There are on average 8.56 million entries in each month of the TRS data. Post-processing, the number of observations is on average 6.16 million per month.

(a) Time stamp adjustment

The firms reporting to TRS operate in different time zones and the data base does not have a built in functionality to adjust the trade times to a common time zone. To adjust the time stamps we record in which hour the first and last exchange-transaction is executed at NASDAQ OMX Stockholm or at one of the MTFs for each firm in each day. All the exchanges open at 8 am (GMT) and close at 4:30 pm (GMT). We adjust the time stamps of firms that do not have their median first and last trades in sync with the opening hours. For example, for a firm with the median first trade hour in a month being 7 am and the median last trade hour being 3 pm, we adjust all transaction time stamps by +1 hour.

(b) Matching to TRTH data

TRS transactions are matched to TRTH transactions using information on stock, trading venue, date, time, price, buy/sell, and quantity. The time stamps of the two databases are allowed to differ by no more than one second. Where trader IDs are available in the TRTH data (for NASDAQ OMX Stockholm only), they are added to the matching criterion. TRTH trades are split into two transactions, one for the buyer and one for the seller. If there are several matches to one transaction, the transaction closest in time is considered to be the closest match. To the extent that there are multiple TRTH trades in the same second, same stock, at the same trading venue, with the same price and quantity, and with different sub-second time stamps, this approach may introduce noise in the time stamps. This potential problem applies only where trader IDs are unavailable in TRTH, and it is more likely at the most active trading venue, for example December 2014 at NASDAQ OMX Stockholm. For that subset, 6.5 % of all transactions have a potentially noisy sub-second time stamp. The time stamp noise due to this problem is however unlikely to influence the *Decision Latency*, see discussion in footnote 16.

(c) Firms

We analyze trading revenues at the corporation level rather than the branch or division level. Accordingly, we truncate BIC codes (11-letter identifiers of the financial institutions reporting to TRS) to the first four letters that are unique to each corporation. For various reasons, such as mergers and acquisitions, the same corporation may span several (truncated) BIC codes. For example, GETCO acquired

Knight Trading in August 2013. We thus treat the (truncated) BIC codes GEEU (GETCO) and NITE (Knight Trading) as separate for the period preceding the merger and as one corporation for August 2013 onwards.

(d) Filtering of trades

We exclude transactions where the trade price is more than 5% lower than the official low price of the stock-day, or more than 5% higher than the official high price of the day. The official statistics do not consider OTC transactions, so prices outside the High-Low interval are possible, but deviations of more than 5% are considered erroneous. TRS transactions that are flagged as derivative-related either in TRS or in the TRTH entry that it is matched to are also excluded.

Non-proprietary transactions frequently generate more than one entry in TRS. For example, if a broker buys 100 shares on behalf of a client, it may be reported as two transactions in TRS: one transaction where the reporting firm purchases 100 shares at the exchange, and one off-exchange transaction where the reporting firm sells 100 shares to its client.[29] As firms differ in how they report their transactions we need to process the data to make transactions comparable.

For each transaction we seek to retain one representative TRS entry and to attach an entry of the *end investor* associated with that transaction. The *end investor* assignment is done differently depending on the type of trade.

We define *primary transaction* as TRS entries where the counterparty of the trade is a clearing house or the same as the trading venue for the transaction **or** the owner of the trading venue (in the case of dark pools). The definition is motivated by the fact that all exchange transactions must be done through central counterparty (CCP) clearing. All other TRS entries are defined as *secondary transactions*. Of all TRS entries, 81.5% are considered primary transactions.

(i) Primary transaction matched to a secondary transaction of the same firm

To account for several entries reported for the same transaction, we match primary and secondary transactions by *reporting firm*, *stock*, *price*, *quantity*, *date*, and *time*. The time stamps are allowed to differ by no more than one second. The end investor of primary trades matched in this way is set to the client of

[29] In a memorandum on transaction reporting FI provides numerous examples on how different types of trades on behalf of clients may be reported. The memo may be retrieved at http://www.fi.se/upload/90_English/90_Reporting/TRS/memo_transaction_reporting_ver_1_7_2014-03-07.pdf

the secondary trade, if available, and otherwise to the counterparty of the secondary trade. The matched secondary trades are then discarded. Of all primary transactions, 26 % are matched to a secondary transaction in this way.

(ii) Primary transaction matched to a secondary transaction of another firm

To account for the case that the same transaction is reported by both counterparties, we match the reporting firm of primary transactions to the counterparty of secondary transactions. The other matching criteria include *stock*, *price*, *quantity*, *date*, and *time*. As above, the time stamps are allowed to differ by no more than one second. The end investor of primary trades matched in this way is set to the client of the secondary trade, if available, and otherwise to the *reporting firm* of the secondary trade. The matched secondary trades are then discarded. Of all primary transactions, 12 % are matched to a secondary transaction in this way.

(iii) Primary transaction that is not matched to a secondary transaction

Primary transactions that are not matched to a secondary transaction are considered to be on behalf of a client if a client reference is available, and otherwise proprietary. For client (proprietary) trades, the end investor is set equal to the client reference (the reporting firm).

(iv) Secondary transaction that is not matched to a secondary transaction

Secondary transactions that are not matched to a primary transaction, are considered to be on behalf of a client if a client reference is available, and otherwise proprietary. For client (proprietary) trades, the end investor is set equal to the client reference (the reporting firm).

(v) Secondary transactions where the counterparty does not report to TRS

To capture firms that are not obliged to report to TRS, but that still trade our sample stocks, we look for firms that are reported as counterparties but that not show up as reporting firms. For all secondary transactions where such firms appear as counterparties, we create a new entry with the same properties but with opposite direction of trade and with counterparties reversed. This is a way to detect HFT firms that connect to the market through direct market access or sponsored access. Of all secondary transactions, 16 % are subject to this procedure.

A2. HFTs active at NASDAQ OMX Stockholm

Appendix Table 1 reports 19 firms identified as HFTs in the sample. Due to confidentiality requirements, we cannot report the full list of names of the 25 HFTs covered in the proprietary data set. However, in Appendix Table 1, we use public trading records to report the names of 19 HFTs who trade at NASDAQ OMX Stockholm as members. The HFTs not listed in Appendix Table 1 therefore trade only at other trading venues or as clients of other members at NASDAQ OMX Stockholm.

INSERT APPENDIX TABLE 1 ABOUT HERE

A3. Comparison of HFT revenue calculation methods

We compare four methods of calculating trading revenues. As explained in the main text, adjustments are needed because small data errors in inventory can accumulate over time, leading to large and persistent (unit root) errors in computing position that can persist indefinitely throughout the sample if left uncorrected. However, this does not appear to be much of a concern in practice, as shown below (see Appendix Table 2). The four methods are as follows. *No adjustment* is calculated by cumulating daily inventory positions over the full sample. *Method 1: Benchmark* is the method used throughout the paper that zeros the end-of-day position daily for each HFT firm (equivalent to assuming that each firm liquidates any remaining end-of-day position at the daily closing price). *Method 2: Intraday Revenues* assumes a first-in-last-out inventory accounting. That is that any remaining end-of-day positions were never purchased in the first place. *Method 3: Intraday Revenues Plus Revenues from Inventory Sold* is similar to *Method 2* but adds back in the revenues from closing end-of-day positions that are in opposite direction of previous day end-of-day inventory. That is, the end-of-day inventory is marked to market only if an offsetting position exists in the previous end-of-day inventory.

Appendix Table 2 reports the firm cross-sectional distribution of HFT trading revenues using the four different methods for calculating trading revenues. It shows that inventory adjustments do not alter the main results of the paper in the sense that the mean, median, and distribution are roughly similar across the different methods. As a result, marking-to-market end-of-day inventory positions is relatively innocuous.

INSERT APPENDIX TABLE 2 ABOUT HERE

A4. Construction of daily Fama-French plus momentum factors for Swedish equities

We construct daily Fama-French and momentum factors for Swedish equities. The dataset used to construct the factors (using the variables: daily total excess returns, shares outstanding, and quarterly book values) comes

from Compustat Global and covers the period January 2010 to December 2014. We exclude stock-day observations in which the total market capitalization falls below 100 MSEK (about 10.5 million USD as of the exchange rate of December 2014). The four factors (excess market return, small minus large [SML], value minus growth [HML], winner minus loser [WML]) are constructed according to the specifications used to create U.S. factors, as specified on Kenneth French's website: the value-weighted portfolios consist of top-30%, middle 40%, and bottom-30% of stocks (by market capitalization, book to market, and past-12-month returns for SML, HML and WML, respectively) and are re-sorted every July 1 using data from the previous year's performance.

Appendix Table 3 reports summary statistics corresponding to these traded risk factors for Swedish equities. The statistics include the mean daily log excess return (annualized), its standard error, and the number of observations (i.e., the number of trading days), and are reported for each portfolio. The annualized excess returns on the four portfolios (market excess return, SMB, HML, WML) are 0.160, 0.176, 0.039, and 0.028, respectively, which are all positive, as expected.

INSERT APPENDIX TABLE 3 ABOUT HERE

A5. Exchange fees, liquidity rebates, and their potential effect on trading performance

Appendix Table 4 reports exchange fees in 2014 for three stock exchanges (NASDAQ OMX Stockholm, BATS, and Chi-X) trading Swedish equities. Fees range from 0 to 0.325 bps over these selected venues. Exchange fees depend on the side of the trade: "maker" fees are less than "taker" fees (for example, 0.13 vs. 0.325 bps for NASDAQ OMX S30 stocks), and, at Chi-X, makers receive liquidity rebates (negative fees) of about 0.225 bps. For NASDAQ OMX Stockholm, we report the fees for S30 stocks, which are lower than for other stocks; all the stocks in our sample fall into this category.

INSERT APPENDIX TABLE 4 ABOUT HERE

While NASDAQ OMX Stockholm grants preferential prices for liquidity provision under its Liquidity Provider Scheme (LPS), BATS and Chi-X do not (a designated liquidity provider program exists but doesn't have lower fees). Although BATS and Chi-X merged in November 2011, with technology integration complete by April 2012, the trading platforms continue to implement different pricing structures.

Appendix Table 5 analyzes HFT performance after accounting for potential maker-taker fees and liquidity rebates and shows that even accounting for the most conservative possible fees and/or rebates does not qualitatively change the results.

INSERT APPENDIX TABLE 5 ABOUT HERE

This table is similar to Table 2 but adjusts for potential maker-taker fees and liquidity rebates. Panel A reports trading revenues under the assumption of the maximum possible maker-taker fees on NASDAQ OMX Stockholm (0.325 bps taker fees; 0.125 bps maker fees), and Panel B uses the maximum possible on the Chi-X exchange, which features a liquidity rebate (0.30 bps taker fees; 0.225 bps liquidity rebate).

Accounting for the most conservative possible fees and/or rebates does not qualitatively change the results. For example, though trading performance for the entire distribution is shifted down slightly, the sign of *Revenues* does not change for any HFT firm. We additionally confirm (not reported in the table) that the performance results are still positively skewed, with the same HFTs at the top strongly outperforming their competitors.

A6. Trading performance and latency of the 5 fastest HFTs

Appendix Table 6 is similar to Table 4 but breaks down the *Top 1-5* dummy variables into individual dummy variables for the fastest HFTs: *Top 1*, *Top 2*, *Top 3*, *Top 4*, and *Top 5*. It shows that even among the top five HFTs the faster firm tends to perform better and that performance is monotonic in latency.

INSERT APPENDIX TABLE 6 ABOUT HERE

Appendix Table 1: HFTs active at NASDAQ OMX Stockholm

This table shows a list of HFTs who according to public trading records are active at NASDAQ OMX Stockholm for at least one month of our sample period (January 2010 to December 2014). The list contains 19 of the 25 firms identified as HFTs in our sample. We can only show the public record for confidentiality reasons. HFTs that are not listed do not trade as members at NASDAQ OMX Stockholm, but may trade at other trading venues or as clients of other members at NASDAQ OMX Stockholm. The HFTs are presented in alphabetical order.

Algoengineering
All Options International
Citadel Securities
Flow Traders
GETCO[a]
Hardcastle Trading
IMC Trading
International Algorithmic Trading (SSW Trading)
Knight Capital [a]
Madison Tyler[b]
MMX Trading
Optiver
Spire
Susquehanna Int. Sec.
Timber Hill
WEBB Traders
Virtu Financial[b]
Wolverine Trading UK

[a] Knight Capital merged with GETCO in July 2013 (https://www.sec.gov/Archives/edgar/data/1569391/000119312513279128/d559202d8k12g3.htm)
[b] Madison Tyler merged with Virtu Financial in July 2011 (https://www.sec.gov/Archives/edgar/data/1592386/000104746914002070/a2218589zs-1.htm)

Appendix Table 2: Comparison of HFT revenue calculation methods

This table reports the firm cross-sectional distribution of HFT trading revenues (as in Table 2) using four different methods for calculating trading revenues. *No adjustments* is calculated by cumulating daily inventory positions over the full sample; *Method 1: Benchmark* is the method used throughout the paper that zeros the end-of-day position daily for each HFT firm (equivalent to assuming that each firm liquidates any remaining end-of-day position at the daily closing price); *Method 2: Intraday Revenues* assumes that any remaining end-of-day positions were never purchased in the first place (assuming first-in-last-out inventory accounting); *Method 3: Intraday Revenues Plus Revenues from Inventory Sold* is similar to *Method 2* but adds back in the revenues from closing end-of-day positions that are in opposite direction of previous day end-of-day inventory (that is, the end-of-day inventory is marked to market only if an offsetting position exists in the previous end-of-day inventory).

	mean	stdev	p10	p25	p50	p75	p90
No adjustments							
Revenues	20,824	25,149	-6,788	277	10,037	45,082	65,415
Sharpe Ratio	4.07	5.16	-0.76	0.80	2.81	6.04	10.89
Revenues per MSEK Traded	159.42	326.12	-201.36	27.00	59.05	217.77	443.20
Method 1: Benchmark							
Revenues	18,181	29,519	-7,572	-487	6,990	31,968	61,354
Sharpe Ratio	4.16	6.58	-1.47	0.33	1.61	7.02	11.14
Revenues per MSEK Traded	153.25	504.78	-257.94	-43.71	56.45	147.24	472.16
Method 2: Intraday Revenues							
Revenues	18,069	29,527	-7,554	-577	7,095	31,972	61,243
Sharpe Ratio	4.15	6.50	-1.47	0.33	1.61	7.01	11.14
Revenues per MSEK Traded	146.98	476.65	-255.17	-24.76	55.54	147.03	469.27
Method 3: Intraday Revenues Plus Revenues from Inventory Sold							
Revenues	21,128	25,451	-2,036	2,026	11,408	32,193	66,835
Sharpe Ratio	3.34	3.80	-0.09	0.75	1.93	4.90	9.37
Revenues per MSEK Traded	265.14	535.64	-200.68	9.13	88.62	367.66	1,160.62

Appendix Table 3: Daily Fama-French plus momentum factors for Swedish equities

This table reports summary statistics corresponding to daily Fama-French plus momentum factors created for Swedish equities. The mean daily log excess return (annualized), its standard error, and the number of observations (i.e., number of normal trading days) are reported for each of the portfolios. The four factors are constructed according to the specifications used to create U.S. factors, as specified on Kenneth French's website: the value-weighted portfolios consist of top-30%, middle 40%, and bottom-30% of stocks and are re-sorted every July 1 using data from the previous year's performance. The data (daily total excess returns, shares outstanding, and quarterly book values) come from Compustat Global and covers the period January 2010 to December 2014.

	Mean	S.E.	Daily observations
log market excess returns	0.160	0.083	1255
log large-cap returns	0.152	0.085	1255
log medium-cap returns	0.231	0.070	1255
log small-cap returns	0.347	0.063	1255
log SML returns	0.176	0.068	1255
log growth returns	0.161	0.088	1255
log neutral returns	0.143	0.084	1255
log value returns	0.206	0.090	1255
log HML returns	0.039	0.054	1255
log winner returns	0.171	0.095	1255
log neutral returns	0.155	0.084	1255
log loser returns	0.136	0.088	1255
log WML returns	0.028	0.065	1255

Appendix Table 4: Exchange fees for three exchanges trading Swedish equities

This table reports exchange fees in 2014 for three stock exchanges (NASDAQ OMX Stockholm, BATS, and Chi-X) trading Swedish equities. Exchange fees depend on the side of the trade: "maker" fees are less than "taker" fees, and, at Chi-X, makers receive liquidity rebates (negative fees). NASDAQ OMX Stockholm fees are lower for S30 stocks; all the stocks in this study fall into this category. While NASDAQ OMX Stockholm grants preferential prices for liquidity provision under its Liquidity Provider Scheme (LPS), BATS and Chi-X do not (a designated liquidity provider program exists but it does not have lower fees). Although BATS and Chi-X merged in November 2011, with technology integration complete by April 2012, the trading platforms continue to implement different pricing structures.

	NASDAQ OMX Stockholm for S30 stocks	NASDAQ OMX Stockholm Liquidity Provider Scheme (LPS) for S30 stocks	BATS*	Chi-X*
Maker	0.125 bps	0 bps	0 bps	-0.15 to -0.225 bps**
Taker	0.325 bps	0.5 bps	0.15 bps	0.30 to 0.24 bps

* For non-hidden limit orders

** The exact price within this range depends on volume. The lowest fees are given after total monthly trading volume exceeds 16 billion EUR. Negative values represent liquidity rebates.

Appendix Table 5: HFT performance after accounting for potential maker-taker fees

This table is similar to Table 2 but adjusts for potential maker-taker fees and liquidity rebates. Panel A reports trading revenues under the assumption of the maximum possible maker-taker fees on NASDAQ OMX Stockholm (0.325 bps taker fees; 0.125 bps maker fees), and Panel B uses the maximum possible on the Chi-X exchange, which features a liquidity rebate (0.30 bps taker fees; 0.225 bps liquidity rebate).

Panel A: Using maker-taker fees on NASDAQ OMX Stockholm

	Mean	Std.Dev.	p10	p25	p50	p75	p90
Revenues (SEK)	12,012	18,688	-2,228	-1,212	5,124	19,346	52,553
Revenues per MSEK Traded	31.87	207.36	-209.56	-39.64	34.68	66.04	330.17
Returns	0.22	0.33	-0.08	0.00	0.08	0.45	0.81
Sharpe Ratio	3.40	5.41	-1.50	0.00	1.66	6.20	10.49
1-factor Alpha	0.22	0.33	-0.07	0.00	0.08	0.45	0.82
3-factor Alpha	0.22	0.33	-0.06	0.01	0.08	0.45	0.80
4-factor Alpha	0.22	0.33	-0.06	0.01	0.08	0.45	0.80

(N = 16 firms)

Panel B: Using taker fees & liquidity rebates on Chi-X

	Mean	Std.Dev.	p10	p25	p50	p75	p90
Revenues (SEK)	17,525	24,207	-1,230	-327	5,252	36,366	56,420
Revenues per MSEK Traded	50.41	208.01	-198.67	-22.39	59.28	83.61	362.39
Returns	0.34	0.52	-0.08	0.01	0.08	0.48	1.11
Sharpe Ratio	4.74	7.61	-1.45	0.40	1.69	6.56	13.53
1-factor Alpha	0.34	0.53	-0.07	0.01	0.08	0.48	1.12
3-factor Alpha	0.34	0.54	-0.06	0.02	0.09	0.48	1.11
4-factor Alpha	0.34	0.54	-0.06	0.02	0.09	0.48	1.11

(N = 16 firms)

Appendix Table 6: Trading performance and latency of the 5 fastest HFTs

This table is similar to Table 4 but breaks down the *Top 1-5* dummy variables into individual dummy variables for the fastest HFTs: *Top 1*, *Top 2*, *Top 3*, *Top 4*, and *Top 5*. As in Table 4, it reports coefficients estimated from Eq. (1) for five performance measures as dependent variables: *Revenues*, *Returns*, *Sharpe Ratio*, *Trading Volume*, and *Revenues per MSEK Traded* (all defined as in Table 2). We estimate OLS regressions with month fixed effects. The independent variables considered are as follows: log(*Decision Latency*) is the natural logarithm of *Decision Latency* (defined as in Table 2). *Top 1*, *Top 2*, *Top 3*, *Top 4*, and *Top 5* are indicator variables for whether a given firm is ranked as the top 1, 2, 3, 4, or 5 firms by *Decision Latency* in a given month. The control variables, whose estimated coefficients are omitted to conserve space, are the same as in Table 4. All continuous independent variables are in units of standard deviations. *, ** and *** correspond to statistical significance at the 10%, 5%, and 1%, respectively. Standard errors are dually clustered by firm and month and are reported in the parentheses. The sample consists of 25 Swedish stocks and 60 months of trading (January 2010 to December 2014).

	Revenues			Returns			Sharpe Ratio			Trading Volume (x 10^{-6})			Revenues per MSEK Traded		
log(Decision Latency)	-14020***	1275	11334	-.221***	-.0297	.0174	-4.38***	-.395	2.34	-247***	-54.5	32.6	-19.4	-6.68	102**
	(4311)	(6837)	(10708)	(.0483)	(.063)	(.0797)	(.632)	(1.34)	(1.5)	(43.7)	(64.7)	(77.8)	(57.5)	(69.3)	(40.6)
Top 1		58288***	44007***		.626***	.604***		12.2***	10.8***		693***	545***		33.9	106*
		(15035)	(11886)		(.152)	(.166)		(3.63)	(3.7)		(160)	(145)		(103)	(57.4)
Top 2		40358***	29165***		.396***	.361***		9.01**	6.98**		496***	371***		26.7	56.3
		(15163)	(10847)		(.15)	(.134)		(3.55)	(3.12)		(143)	(101)		(116)	(83.8)
Top 3		35262*	24967*		.575***	.534***		12.1**	9.91**		563**	446***		23.8	33.9
		(21206)	(14052)		(.212)	(.19)		(5.17)	(4.22)		(232)	(169)		(98.2)	(63.2)
Top 4		23756**	16243**		.441**	.414**		8.89**	7.01**		249*	159		78.3	90.5***
		(11854)	(7722)		(.21)	(.182)		(3.51)	(2.88)		(136)	(108)		(88)	(33.5)
Top 5		11588	5422		.106	.0955		2.92	2.21		117	48.6		-25.9	14.8
		(7457)	(7618)		(.0952)	(.0838)		(1.8)	(1.62)		(87.5)	(79)		(102)	(73.7)
Constant	20278***	6993*	9568**	.254***	.0852	.0885*	5.1***	1.55	1.92	313***	147***	177***	35.2	24.4	6.17
	(6973)	(3874)	(4393)	(.0579)	(.0564)	(.0463)	(1.26)	(1.28)	(1.21)	(75.9)	(55.8)	(52.4)	(57.3)	(82.2)	(11.9)
Controls	No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes
Month FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	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