Need for Speed? Exchange Latency and Liquidity

Albert Menkveld, Marius Andrei Zoican

Duisenberg school of finance - Tinbergen Institute Discussion Paper

TI 14-097/IV//DSF78

Need for Speed? Exchange Latency and Liquidity

Albert J. Menkveld
Marius A. Zoican

Faculty of Economics and Business Administration, VU University Amsterdam, Duisenberg School of Finance, and Tinbergen Institute.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Need for Speed? Exchange Latency and Liquidity

Albert J. Menkveld and Marius A. Zoican*

July 22, 2014

Abstract

Speeding up the exchange does not necessarily improve liquidity. The price quotes of high-frequency market makers are more likely to meet speculative high-frequency "bandits," thus less likely to meet liquidity traders. The bid-ask spread is raised in response. The recursive dynamic model reveals that there is an additional spread-widening effect as market makers earn higher rents due to economies of scope from quote monitoring. Analysis of a NASDAQ-OMX speed upgrade provides supportive evidence.

Keywords: market microstructure, trading speed, information asymmetry, high-frequency trading

JEL Codes: G11, G12, G14

*Both authors are affiliated with VU University Amsterdam, the Tinbergen Institute, and Duisenberg School of Finance. Address: FEWEB, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands. Albert Menkveld can be contacted at albertjmenkveld@gmail.com. Marius Zoican can be contacted at m.a.zoican@vu.nl. The paper received the "Outstanding Paper in Investments" award at the Eastern Finance Association 2014. We have greatly benefited from discussion on this research with Istvan Barra, Alejandro Bernales, Jonathan Brogaard, Eric Budish, Sabrina Buti, Lucyna Górnicka, Terrence Hendershott, Søren Hesel, Ali Hortaçsu, Bernard Hosman, Peter Hoffmann, Johan Hombert, Wenqian Huang, Olga Lebedeva, Sophie Moinas, Emiliano Pagnotta, Talis Putnins, Ryan Riordan, Satchit Sagade, Vincent van Kervel, and Bart Yueshen Zhou. Albert Menkveld gratefully acknowledges NWO for a VIDI grant. We are grateful to the participants from the Tinbergen Institute and DSF Brown Bag Seminars for insightful comments, as well as conference participants at University College London, SGF Zürich, ERIC Stuttgart Doctoral Consortium, the Center for Financial Studies in Frankfurt, the 26th Australasian Banking and Finance Conference, the 6th Hedge Fund Research Conference in Paris, the 2014 Eastern Finance Association Meeting, the 2014 FIRS meeting, and the 2014 Midway Market Design Workshop at the University of Chicago.

Need for Speed? Exchange Latency and Liquidity

Abstract

Speeding up the exchange does not necessarily improve liquidity. The price quotes of high-frequency market makers are more likely to meet speculative high-frequency "bandits," thus less likely to meet liquidity traders. The bid-ask spread is raised in response. The recursive dynamic model reveals that there is an additional spread-widening effect as market makers earn higher rents due to economies of scope from quote monitoring. Analysis of a NASDAQ-OMX speed upgrade provides supportive evidence.

Keywords: market microstructure, trading speed, information asymmetry, high-frequency trading
JEL Codes: G11, G12, G14

1 Introduction

> It reminds me of the old story of the two high-frequency traders on safari. Coming out of the jungle into a clearing, they are faced with a hungry lion, staring at them and licking his lips. One of the traders immediately starts taking off his boots and donning a pair of sneakers. "What are you doing?" says the other trader. "You'll never be able to outrun a hungry lion." "I don't need to outrun the lion," says the first trader. "I only need to outrun you."
>
> — HFT Review, April 2010

Speed matters for individual agents, but does it matter for an exchange? Do markets get better when trading platforms reduce their latency? Pagnotta and Philippon (2013) document the many speed investments that exchanges around the world have implemented between 2008 and 2012. For instance, the trading latency on the New York Stock Exchange (NYSE) dropped from 350 milliseconds in 2007 to 5 milliseconds in 2009. The industry started referring to the speed of light as a binding constraint.[1]

A speed improvement at modern exchanges only directly affects those who employ computerized trading strategies. Budish, Cramton, and Shim (2013) consider human reaction time to be hundreds of milliseconds. High-frequency traders (HFTs) are at the other extreme. They are characterized as proprietary traders who "use of extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders (SEC, 2010, p. 45)."

The impact of exchange latency depends on the strategies employed by the ultra-fast traders, i.e., HFTs. They act both as market makers and as short term speculators (SEC, 2010). A key difference between the two types is that the former predominantly submits price quotes (limit orders) whereas the latter mostly consumes price quotes (market orders). Hagstromer and Norden (2013) document that such order type specialization exists for high-frequency traders (HFTs) at NASDAQ-OMX. Baron, Brogaard, and Kirilenko (2012) find a similar specialization for HFTs at the Chicago Mercantile Exchange.

This paper analyzes the impact of exchange latency on liquidity in the presence of two canonical HFT strategies. On the one hand, a faster market allows high-frequency market makers (HFMs) to update their

[1] See How Low Can You Go?, HFT Review, April 2010.

price quotes faster on new (public) information. On the other hand, it enables high-frequency speculators, referred to as bandits (HFBs), to act faster on this new information and profit by trading against potentially stale quotes. Others, referred to as liquidity traders, are assumed to be uninformed and not directly affected by an exchange speed creeping closer to the speed of light.

There is no exogenous information asymmetry between market makers and speculators. Adverse selection for HFMs is generated by equally fast HFBs. It is a game of chance who is first to the market after news. This type of adverse selection risk complements the existing literature that focuses on exogenous information asymmetry as a source of adverse selection risk (e.g., Glosten and Milgrom, 1985; Kyle, 1985; Foucault, Hombert, and Roşu, 2013). Adverse selection on news is also part of Budish, Cramton, and Shim (2013) who focus on technology investments by high-frequency traders. Foucault, Kozhan, and Tham (2014) use it to study "toxic arbitrage."

The paper's main result is that lowering exchange latency (i.e., increasing speed) reduces liquidity. We identify two main channels that drive this result: a static and a dynamic one. A faster market essentially makes the trading game in each (shorter) interval more of a game between markets makers and bandits, HFMs and HFBs. The arrival rate of liquidity traders remains unchanged and their participation in each trading game therefore declines. Competitive HFMs have to raise the bid-ask spread to recoup their increased loss due to more trades with HFBs. Liquidity traders suffer as they pay the higher spread. We label this the static channel.

A faster market further reduces liquidity through a dynamic channel. A market maker has an incentive to update his outstanding quotes on incoming news so as to avoid being adversely selected. Parsing all relevant news instantaneously requires costly computing power as it involves vast amounts of information, e.g., any information sent through newswire services, order book activity, same-industry stock activity, index activity, etc. This costly monitoring however can be amortized across all quotes the market maker has outstanding for the security. This amortization creates an opportunity for rents when the competitive threat for an "incumbent" HFM comes from an entrant HFM with no presence in the order book yet. In the stage game that starts with a book that has an opportunity to refill as a liquidity trader just consumed a quote, the incumbent HFM optimally quotes at the reservation price of the entrant HFM. This reservation price exceeds his own due to his

amortization benefit. In this stage game the incumbent HFM therefore makes a positive profit in expectation.

When the exchange speeds up, the stage game that enables an incumbent HFM to earn rents becomes more likely. This is the dynamic channel through which liquidity is reduced. The intuition for this result is that this stage game only materializes if in the previous time interval a liquidity trader arrived, *and* there was no news. It is the latter event that is the engine of the result. Had there been news as well, then all book quotes effectively become stale and useless. The HFMs compete again on both sides of the book as it needs to fill up around the new fundamental value. The incumbent HFM advantage no longer exists. As the market speeds up, the probability of two events (liquidity trader arrival and news) in an inter-HFT-arrival interval declines as this interval becomes shorter. The steady state probability of only a liquidity trader arrival increases, and the recursive equilibrium therefore implies more rents for HFMs on average, effectively paid for by liquidity traders through a larger spread. Liquidity deteriorates.

The paper generates two more results. First, faster markets imply more quote flickering. This is caused directly by transiting between the zero-rent and positive-rent stage games. The spread in the latter game is higher. The flickering result is there absent strategic order submissions and cancelations, a channel explored in Pagnotta and Philippon (2013), Baron, Brogaard, and Kirilenko (2012), and Yueshen (2014). Second, beyond a threshold speed HFMs start paying a monitoring cost to protect themselves against HFBs. This investment in and of itself is a social cost as improved monitoring pays off privately but not publicly (as liquidity traders are uninformed). The baseline model assumes that speed is beyond the threshold level. The aforementioned dynamic liquidity channel and quote flickering critically depend on active monitoring.

A NASDAQ-OMX system speed upgrade is used to empirically test the implications of the model. On February 8, 2010, INET Core Technology was introduced for equity trading in Denmark, Finland, and Sweden. Exchange latency dropped from 2.5 to 0.25 milliseconds and colocation was introduced; both significantly boosted HFT speed. Following this upgrade, adverse selection cost for high-frequency market makers increased significantly by 2.11 basis points. This change is large economically as it implies that the post-event level is more than five times higher than the pre-event level of 0.39 basis points (standard control variables for bid-ask spread analysis were included). The effective spread charged by them increased by 32% as the adverse selection cost increase is partially offset by a lower HFM realized spread (i.e., their gross

profit). This spread decomposition result is there also for quote submitters other than the identified HFMs, albeit less pronounced.

A calibration of the model reveals that the huge increase in adverse selection cost for HFMs can be achieved for reasonable parameter values. It relies strongly on the stylized fact that trades cluster in time. A sub-second speed change has substantial bite when securities markets feature short bursts of activity.

In sum, the paper's message is that further speeding up an extremely fast exchange could hurt liquidity, *ceteris paribus*. This is somewhat surprising given that a large literature has emerged that by and large finds that liquidity benefits from migrating from human-intermediated trading to algorithmic trading (see Section 2). A first-order effect is that automation tends to reduce cost in any industry. The larger point of this paper is that once automation is in place, speeding up further is not necessarily good in the securities trading industry (unlike other industries, e.g., smart phones, online backup systems, gaming, etc.).[2]

The rest of the paper is structured as follows. Section 2 briefly reviews the literature on the advantages and disadvantage of low-latency trading and positions the paper in this literature. Section 3 develops a recursive trading model with competitive high-frequency market makers and high-frequency speculators. Section 4 exploits a natural experiment to test the empirical predictions of the model. The model is calibrated to the data in Section 5. Section 6 concludes.

2 Related literature

High-frequency trades and liquidity. This paper is part of a rapidly growing literature on high-frequency traders (HFTs) and liquidity. Foucault, Hombert, and Roşu (2013) argue that HFTs have better information as they can process news faster. Information asymmetry is increased as a result. Similarly, Martinez and Roşu (2013) argue that the informational advantage of HFTs increases volatility and reduces liquidity, in a model where HFTs pick off competitive market makers' quotes. Hoffmann (2013b) shows that while higher speed generates positive gains from trade, the bargaining power shifts from slow to fast traders. The latter capture

[2] Budish, Cramton, and Shim (2013) note that in the last decade the gradual speed increase coincided with lower duration for arbitrage opportunities (their own work), a lower bid-ask spread (Virtu IPO filing), and a lower cost of trading large blcoks (Angel, Harris, and Spatt, 2013). More importantly, they note that all these trends seem to flatten out. Our contribution is to show that liquidity could actually deteriorate at an extreme speed.

the full surplus. Biais, Foucault, and Moinas (2013) propose a model of low-latency trading where when a subset of agents become fast, all other traders incur higher adverse selection cost. Jovanovic and Menkveld (2011) argue that HFT entry has an ambiguous effect on welfare. If HFTs are the only agents with access to information, they can reduce gains from trade. Aït-Sahalia and Saglam (2014) propose a dynamic trading model of an HFT market maker who receives a signal about future order flow. They consider the effect of various regulatory policies. Jovanovic and Menkveld (2014) find that if there is a (small) participation cost, then the availability of more HFTs widens the bid-ask spread.

Several empirical papers suggest a positive relationship between the adverse selection cost incurred on limit orders and high-frequency trading. Brogaard, Hendershott, and Riordan (2014) and Hendershott and Moulton (2011) document a larger permanent price impact for market orders when they are sent by HFTs. In the same line, Baron, Brogaard, and Kirilenko (2012) show that HFTs earn short-term profits on the market orders they submit, consistent with them adversely selecting others. Hoffmann (2013a) focuses on adverse selection differentials across market venues. He finds that the adverse selection component is larger on entrant venues. Menkveld (2014) argues that these entrant venues are likely to exhibit high HFT participation. Menkveld (2013) provides supportive evidence as he documents high HFT participation for Chi-X, an entrant venue in Europe. Moallemi and Saglam (2013) estimate the "cost of latency" through the losses from trading on stale information. They find that it amounts to half of total trading costs. Breckenfelder (2013) documents that if more HFTs compete for trades their liquidity consumption increases.

Finally, Jones (2013) and Biais and Foucault (2014) offer a comprehensive review of the theoretical and empirical literature on high-frequency trading and liquidity. They stress that the evidence suggests HFT heterogeneity.

Exchange latency and liquidity. Pagnotta and Philippon (2013) relate exchange competition to market speed. They argue that exchanges can use speed as an instrument to cater to different clienteles. Fast markets are able to charge a premium to traders with volatile private values, i.e., those who value speed most.

The evidence on how exchange latency impacts liquidity is mixed. Riordan and Storkenmaier (2012) find that a latency reduction implemented at the German stock exchange had no detectable effect on the effective

spread for most sample stocks, it only reduced it for the quartile of smallest stocks. Ye, Yao, and Gai (2013) study NASDAQ data and find that a drop in exchange latency from the microsecond to the nanosecond level increased volatility and reduced market depth. It did not have an effect on the effective spread.

Algorithmic trading and liquidity. The set of algorithmic traders includes HFTs but more generally includes all who use computers to automate the trading process. Hendershott, Jones, and Menkveld (2011) document that algorithmic trading causally reduces the bid-ask spread. Similarly, Malinova, Park, and Riordan (2013) find that after the introduction of a message fee, algorithmic trading is crowded out and the bid-ask spread rises.

This paper's contribution. The model proposed in this paper contributes to this literature in the following ways. First, it focuses on the *interaction* of two types of high-frequency traders, market makers and bandits. It explores how this interaction affects liquidity by studying its effect on the trading costs of a third type of agent, the liquidity trader. The recursive model uncovers effects that remain hidden in static models. The modeling framework is particularly useful to get traction on how exchange speed affects trading. Analytic results are derived in a relatively straightforward way.

The empirical contribution is that the adverse selection cost increases when an exchange moves from millisecond to sub-millisecond speed. Trader-level detail allows us to document that this cost increased most for limit orders from HFTs.

3 Model

3.1 Primitives

This section presents the model's primitives. Extensive motivation for these primitives is left to Subsection 3.2. The complete list of model parameters is presented in Appendix A.

Trading environment. A single risky asset is traded on a limit order market with price and time priority. The order book has limited capacity. It can hold only a single order to buy one unit (bid quote) and one order to sell one unit (ask quote). If an HFM submits a strictly better quote in terms of price, then the better quote will replace the existing quote.

Agents. The risky asset is traded by three types of risk-neutral agents: competitive high-frequency market-makers (HFMs), high-frequency speculators or "bandits" (HFBs), and liquidity traders (LTs). The market makers post only limit orders, whereas the other two trader types post only market orders. The setup is similar to Foucault, Roell, and Sandas (2003).

The high-frequency traders (HFMs and HFBs) only submit and cancel orders at fixed times $k\delta$, $k \in \{1, 2, 3, \ldots\}$, where δ is a measure of the exchange latency. For a lower value of δ, high-frequency traders visit the market more often.

Interarrival events. In each time interval δ between high-frequency trader arrivals, two types of events might occur. The common value of the asset changes with probability $\alpha\delta$ with $\alpha > 0$. A common-value shock represents a *news event*. The size of the *news event* can be either σ ('good' news) or $-\sigma$ ('bad' news).

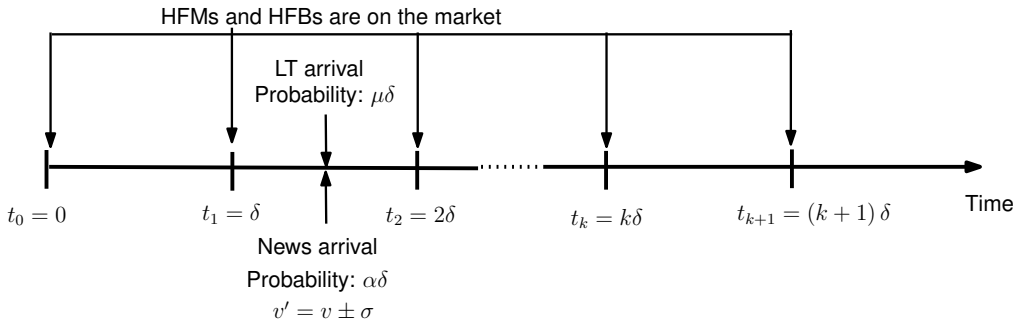
The common value of the asset, v_t, changes from $k\delta$ to $(k+1)\,\delta$ as follows:

$$v_{(k+1)\delta} = \begin{cases} v_{k\delta}, & \text{with probability } 1-\alpha\delta \text{ (no news event)}, \\ v_{k\delta}+\sigma, & \text{with probability } \frac{\alpha\delta}{2} \text{ (good news)}, \\ v_{k\delta}-\sigma, & \text{with probability } \frac{\alpha\delta}{2} \text{ (bad news)}. \end{cases} \tag{1}$$

The other type of event is that an LT might arrive motivated by a private value shock, referred to as a liquidity shock. The probability of this event is $\mu\delta$ with $\mu > 0$. The size of the shock is assumed to be larger than σ.[3] Liquidity shocks are independent from news events. In other words, the common- and private-value shocks are independent.

[3] This assumption is relatively innocent. It rules out market breakdown and keeps the paper's focus on how exchange speed affects liquidity in normal market conditions.

The timing of the trading game is presented below:



Information structure The HFMs can learn the common value by paying a monitoring cost c per time unit. For an interval of length δ this cost is $c\delta$. Monitoring allows them to act on common value innovations and reduce the risk of being picked off by HFBs. High-frequency bandits are fully informed. LTs cannot monitor but infer the common value from HFT activity. Note that LTs are "slow" in the sense that they are they are always the last ones to become informed. They arrive at the market motivated by a private value shock only. They update their belief about the common value from the price quotes they find in the market upon arrival.

Decision ordering HFTs visit the market at $k\delta$, $k \in \{1, 2, 3, \ldots\}$ (and leave the market immediately after the visit). The decisions on market order submission (for HFBs) or limit order cancelation and submission (for HFMs) follow the sequence below:

1. *Order resolution stage.* HFBs decide whether to submit a market order. HFMs decide whether to cancel outstanding limit orders. The market order and the cancellation have the same probability of being executed first.

2. *Monitoring stage.* HFMs decide whether or not to pay the monitoring cost for the next interval.

3. *Order submission stage.* HFMs submit new limit orders at the bid price $v_t - s_{b,t}$ and the ask price $v_t + s_{a,t}$, where $s_{b,t}$ and $s_{a,t}$ are decision variables. HFMs whose quotes were just consumed go first in this stage. These HFMs will be referred to as incumbent HFMs.

3.2 Discussion of the primitives

The trading environment is largely based on Foucault, Roell, and Sandas (2003) and Foucault, Kadan, and Kandel (2013).

High-frequency traders specialize in either market-making (HFMs) or speculation (HFBs). Hagstromer and Norden (2013) and Baron, Brogaard, and Kirilenko (2012) contain evidence in support of such specialization.[4]

Both HFMs and HFBs can monitor the asset value. As in Budish, Cramton, and Shim (2013), latency is the only source of adverse selection. It is the risk that an informed market order executes before the market maker arrives to update his quote. There is no exogenous information asymmetry between market makers and speculators (as in, for example, Glosten and Milgrom, 1985).

Any drop in exchange latency can be exploited only by high-frequency traders as they trade faster than other agents. This is likely to be the case in modern securities markets as they clock in microseconds.

HFMs monitor at higher cost than HFBs. The model can be thought of as a reduced form for the following environment: a "representative" market maker stands in front of many speculators who might observe an inexpensive signal with low frequency. If there are many of these "bandits," then this is equivalent to a single "representative" one who receives such a signal at high-frequency. The market maker needs to acquire all signals that speculators observe to avoid being adversely selected on any particular one.

The monitoring cost increases linearly with the length of the time interval. For algorithms that process information continuously, running for longer time intervals implies a larger cost per time interval. Also, the amount of information to be processed increases with the length of the interval. Even for algorithms not operating continuously, the costs are larger if batches of information to process span a longer period. More information requires more computer resources processing time.[5]

HFMs whose quotes were just consumed are the first to refill the book. Messages on own orders' status arrive

[4]The model's results do not depend on this assumption. A more general model might add an additional round ahead of each stage game and assign HFM and HFB roles randomly across HFTs.

[5]This "real-time" decision making on how to allocate finite computing capacity across securities is different from the *ex ante* decision on how much monitoring technology to install (the latter decision is out of scope here). Several members of industry confirmed that such real-time decisions have become first-order as computing power has become constrained in a world that requires *instantaneously* processing *ever larger* amounts of information (e.g., order book changes, news tickers, activity in correlated securities, etc.).

earlier than public market information (Patterson, Strasburg, and Pleven, 2013). Note that any HFM who refills the book operates under a competitive constraint as other HFMs can undercut his quote immediately after it was submitted. This assumption can easily be relaxed without changing the qualitative predictions of the model.

The current model structure deviates from standard Poisson arrivals in order to get closed-form results. The essential difference is that our model ignores the possibility of two or more arrivals of liquidity traders in a time interval. The same goes for news events. It therefore in a sense really becomes a model for extreme speed only as one can ignore second or higher order terms only if $\delta^2, \delta^3, \ldots$ are small relative to δ. Furthermore, we believe that the model characterizes its main result in a conservative way relative to a Poisson model. We expect the equilibrium effective spread to increase even more in speed when higher order events are considered. The reason is that adverse selection cost scale linearly with the depth of HFM quotes (it takes only one bandit to consume all stale price quotes) whereas it takes multiple liquidity traders to realize a profit on multi-unit depth.

3.3 Solution strategy

The dynamic trading game can be solved as a sequence of static stage games that each depend on the state of the order book. The result is formally stated in Lemma 1.

> **Lemma 1.** (Stage game representation) *The dynamic trading game can be represented by a sequence of static stage trading games. Each stage game (indexed by k) begins with the order submission stage at time* $k\delta$ *and ends with the order resolution stage at decision time* $(k+1)\,\delta$*. The solution of each stage game depends only on the state of the order book at* $k\delta$*: either empty (no quotes), full (quotes on both sides), or half full (a quote on one side of the book only).*

The state space of the model consists of two variables: the value of the asset and the state of the order book at the start of the stage game.

Subsection 3.4 solves for each stage game the HFM monitoring strategy and the equilibrium bid-ask spread, given the state of the order book upon entry of the stage game. Subsection 3.5 considers the effect of latency

on order book state distribution and computes the steady state equilibrium average spread.

3.4 Stage game equilibria

Two stage games are relevant. If the order book is empty, HFMs can post quotes on both sides of the book. If the order book contains one quote, HFMs can submit a limit order on the opposite side. A third case is trivial: if the book is full, HFMs cannot submit new quotes. The first arriving HFM operates under competitive pressure. In what follows, he will be referred to as the HFM. He is aware of rival HFMs who come after him and therefore puts his price quote at a level so that others cannot undercut him profitably. This zero-profit condition for "late" HFMs nails the optimal quote submission strategy of the early HFM. The bid-ask spread that results from this condition is referred to as the competitive spread.

3.4.1 Two-sided price quotes

The competitive half-spread is derived for an HFM who arrives on an empty book and simultaneously posts a quote on both sides of it. The result depends on the HFM's monitoring choice. If the HFM monitors the quotes, his expected profit (π) for the oncoming period is (the subscript $I2$ refers to informed and two-sided quotes)

$$\pi(s_{I2}) = \overbrace{(1-\alpha\delta)}^{\text{No news}}\overbrace{\mu\delta}^{\text{LT}} s_{I2} + \overbrace{\alpha\delta}^{\text{News}}\left[\overbrace{\frac{1}{2}\mu\delta(s_{I2}-\sigma)}^{\text{LT on news side}} + \overbrace{\frac{1}{2}\mu\delta\,(s_{I2}+\sigma)}^{\text{LT on no-news side}} + \overbrace{\left(1-\frac{\mu\delta}{2}\right)\frac{1}{2}\,(s_{I2}-\sigma)}^{\text{No LT and HFB executes}}\right] - c\delta. \quad (2)$$

With probability $(1-\alpha\delta)$ there is no news event. In this case, an LT arrives at the market with probability $\mu\delta$ and HFM earns the half-spread s_{I2}.

With probability $\alpha\delta$ there is a news event and the common value changes. The quote on the innovation side can either be consumed by an LT or an HFB, or cancelled by the HFM. An LT arrives on this side of the market with probability $\frac{\mu\delta}{2}$. If no LT arrives, the HFB arrives before HFM with probability $\frac{1}{2}\left(1-\frac{\mu\delta}{2}\right)$. In both cases, HFM loses $s_{I2}-\sigma<0$. The quote on the opposite side is only consumed by an LT, with probability $\frac{\mu\delta}{2}$. In this case, HFM earns $s_{I2}+\sigma$. Additionally, HFM pays the monitoring cost $c\delta$.

The two-sided competitive half-spread posted by an informed HFM is given by the solution to the zero-profit condition:

$$\pi\left(s_{I2}\right)=0 \Leftrightarrow s_{I2}=\frac{\alpha\sigma\left(2-\mu\delta\right)+4c}{4\mu+\alpha\left(2-\mu\delta\right)}. \tag{3}$$

If the HFM does not monitor the quotes, he never arrives before HFBs. The probability of a trade between HFM and HFB increases from $\frac{1}{2}\left(1-\frac{\mu\delta}{2}\right)$ to $1-\frac{\mu\delta}{2}$. On the other hand, HFM no longer pays the monitoring cost $c\delta$. The profit function for the uninformed HFM is:

$$\pi(s_{U2})=\overbrace{(1-\alpha\delta)}^{\text{No news}}\overbrace{\mu\delta}^{\text{LT}}s_{U2}+\overbrace{\alpha\delta}^{\text{News}}\left[\begin{array}{l}\overbrace{\frac{1}{2}\mu\delta(s_{U2}-\sigma)}^{\text{LT on news side}}+\overbrace{\frac{1}{2}\mu\delta\left(s_{U2}+\sigma\right)}^{\text{LT on no-news side}}+\\ \overbrace{\left(1-\frac{\mu\delta}{2}\right)(s_{U2}-\sigma)}^{\text{No LT and HFB executes}}\end{array}\right]. \tag{4}$$

The two-sided competitive half-spread posted by an uninformed HFM is given by the solution to the zero-profit condition:

$$\pi\left(s_{U2}\right)=0 \Leftrightarrow s_{U2}=\frac{\alpha\sigma\left(2-\mu\delta\right)}{2\mu+\alpha\left(2-\mu\delta\right)}. \tag{5}$$

3.4.2 One-sided price quote

The competitive half-spread is derived for an HFM who arrives on a book that is half full. He posts a price quote on the empty side of the book only. Without loss of generality, we focus on the case where HFM posts an ask quote. The competitive half-spread depends on the strategy of the late HFM (as mentioned earlier in this Subsection). If this HFM monitors his quotes, his expected profit for the oncoming period is:

$$\pi\left(s_{I1}\right)=\overbrace{(1-\alpha\delta)}^{\text{No news}}\frac{\mu\delta}{2}s_{I1}+\overbrace{\alpha\delta}^{\text{News}}\left[\begin{array}{l}\overbrace{\frac{\mu\delta}{2}\left(\frac{1}{2}\left(s_{I1}-\sigma\right)+\frac{1}{2}\left(s_{I1}+\sigma\right)\right)}^{\text{LT on quote side}}+\\ \overbrace{\frac{1}{2}\left(1-\frac{\mu\delta}{2}\right)\frac{1}{2}\left(s_{I1}-\sigma\right)}^{\text{No LT and HFB executes}}\end{array}\right]-c\delta. \tag{6}$$

With probability $(1-\alpha\delta)$ there is no innovation in the common value. With probability $\frac{\mu\delta}{2}$, an LT with a positive private value arrives, i.e., the HFM earns the half-spread s.

With probability $\alpha\delta$ there is a news event. An LT with a positive private value arrives with probability $\frac{\mu\delta}{2}$. If the news is good, the HFM loses $s-\sigma$. If the news is bad, HFM earns $s+\sigma$. With probability $\frac{1}{2}\left(1-\frac{\mu\delta}{2}\right)$, the news is good and no LT arrives in the δ interval, i.e., HFM makes a loss $s-\sigma$ if an HFB order executes first. Additionally, HFM pays the monitoring cost $c\delta$.

The one-sided competitive half-spread posted by an informed HFM is given by the solution to the zero-profit condition:

$$\pi\left(s_{I1}\right)=0 \Leftrightarrow s_{I1}=\frac{\alpha\sigma\left(2-\mu\delta\right)+8c}{4\mu+\alpha\left(2-\mu\delta\right)}. \tag{7}$$

If the HFM does not monitor the quote, he never arrives before HFB. On the other hand, HFM no longer incurs the monitoring cost $c\delta$. The profit function for the uninformed HFM is:

$$\pi\left(s_{U1}\right)=\overbrace{\left(1-\alpha\delta\right)}^{\text{No news}}\frac{\mu\delta}{2}s_{U1}+\overbrace{\alpha\delta}^{\text{News}}\left[\overbrace{\frac{\mu\delta}{2}\left(\frac{1}{2}\left(s_{U1}-\sigma\right)+\frac{1}{2}\left(s_{U1}+\sigma\right)\right)}^{\text{LT on quote side}}+\overbrace{\frac{1}{2}\left(1-\frac{\mu\delta}{2}\right)\left(s_{U1}-\sigma\right)}^{\text{No LT and HFB executes}}\right]. \tag{8}$$

The solution s_{U1} is the same as that for the as for the two-sided order book (see equation (5)). The reason is that only monitoring cost drives a wedge between what an HFM with an outstanding quote might do on the other side of the market relative to one without such outstanding quote. To conserve notation, let

$$s_U \equiv s_{U2}=s_{U1}. \tag{9}$$

Lemma 2 presents the main properties of the competitive half-spreads:

Lemma 2. (Comparative statics) *All competitive half-spreads: s_{I1}, s_{I2}, and s_U, monotonically increase with the size of value innovations (σ) and decrease with the liquidity traders' arrival intensity (μ). Moreover,* $\min\left\{s_{I2}, s_U\right\}$ *and* $\min\left\{s_{I1}, s_U\right\}$ *increase with exchange speed and the news probability (α). And,* $\min\left\{s_{I1}, s_{I2}, s_U\right\}<\sigma$.

3.4.3 Equilibrium strategy for HFM and HFB

The optimal strategy for HFB is to always submit a market order to trade on news. As Lemma 2 finds that the quoted half-spread is smaller than the size of the news, the HFB earns a positive expected profit (in case there are a finite number of them, otherwise their expected profit goes to zero in the limit).

To determine the equilibrium half-spread posted by the HFM, we compare the values of s_{I1}, s_{I2}, and s_U for different values of exchange latency and monitoring cost.

Lemma 3. (Monitoring strategy) *First off,* $s_{I1} > s_{I2}$. *There exist four monitoring cost thresholds,* $\underline{c_1} < \overline{c_1} < \underline{c_2} < \overline{c_2}$, *and two exchange latency thresholds,* $\delta_1 < \delta_2$, *such that:*

(i) ***Full monitoring.*** *HFM monitors both two-sided and one-sided quotes for* $c < \underline{c_1}$ *and any exchange latency, or for* $c \in \left(\underline{c_1}, \overline{c_1}\right)$ *and low exchange latency* ($\delta < \delta_1$).

(ii) ***Partial monitoring.*** *HFM only monitors two-sided quotes for* $c \in \left(\underline{c_1}, \overline{c_1}\right)$ *and high exchange latency* ($\delta \geq \delta_1$), *for* $c \in \left(\overline{c_1}, \underline{c_2}\right)$ *and any exchange latency, or for* $c \in \left(\underline{c_2}, \overline{c_2}\right)$ *and low exchange latency* ($\delta < \delta_2$).

(iii) ***No monitoring.*** *HFM never monitors quotes for* $c \in \left(\underline{c_2}, \overline{c_2}\right)$ *and high exchange latency* ($\delta \geq \delta_2$), *or for* $c > \overline{c_2}$ *and any exchange latency.*

The monitoring strategy defined in Lemma 3 is graphed in Figure 1.

[insert Figure 1 here]

The two-sided informed competitive half-spread is always lower than the one-sided informed competitive half-spread. This reflects the existing economies of scope from monitoring. Information is relatively less expensive if the HFM can share its monitoring cost across multiple quotes. Thus, informed HFMs are able to post narrower spreads by submitting a two-sided order.

For small costs ($c < \underline{c_1}$), the net monitoring benefits are the highest. It follows that $s_{I2} < s_{I1} < s_U$. The

implication is that all quotes are monitored. In equilibrium, an uninformed HFM is undercut by an informed competitor who incurs lower adverse selection cost.

The economies of scope play a role in the optimal monitoring strategy. As costs increase, the HFM stops monitoring one-sided orders first, as the one-sided informed half-spread becomes larger than the uninformed half-spread. If monitoring costs increase even further, the two-sided informed half-spread also exceeds the uninformed half-spread. In this situation, the HFM stops monitoring two-sided quotes as well.

Optimal information acquisition strategies also depend on the exchange speed. As the exchange speed increases, the probability of a trade between the HFM and HFB rises. The expected adverse selection costs for HFM increase as a result. Keeping the monitoring costs constant, a higher speed provides incentives for HFMs to monitor quotes and reduce the risk of having their quotes picked off.

Figure 2 illustrates how the competitive half-spreads s_U, s_{I2}, and s_{I1} change with exchange latency. An intersection point between two curves corresponds to a switch in the monitoring strategy of HFM. For low values of c, only s_U and s_{I1} intersect, i.e., two-sided quotes are always monitored. One-sided quotes are only monitored in fast enough markets. For higher c, only s_U and s_{I2} intersect, i.e., one-sided quotes are never monitored whereas two-sided quotes are only monitored in fast enough markets.[6]

[insert Figure 2 here]

Definition 1. The conditional half-spread is defined as the competitive half-spread. It is conditional in the sense that it depends on the order book. Depending on the order book state, we define a two-sided and a one-sided conditional half-spread.

1. If the HFM can quote on both sides of the market it is $s_2^* = \min\{s_{I2}, s_U\}$.
2. If the HFM can only post on one side of the market it is $s_1^* = \min\{s_{I1}, s_U\}$.

Proposition 1 describes the full equilibrium strategies the stage game.

[6]The parameter values were chosen so as to illustrate the economic forces in the model. They are fixed throughout so that one can compare across figures. We set the intensity of news and liquidity-trader arrival at about the same level. This is consistent with price-deformation models that demonstrate that price changes and transaction rates tend to cluster (Engle and Russell, 1998). Periods of high volatility are likely to increase liquidity trading (e.g., replicating stragies for derivatives) .

Proposition 1. (Stage game equilibrium strategy for HFM and HFB) *The equilibrium strategy in the stage game is as follows:*

(i) The incumbent HFM cancels all limit orders if there was news event in the previous period.

(ii) The incumbent HFM does not cancel any of his outstanding limit orders if there was no news in the previous period.

(iii) An HFM posts a sell price quote at $v_t + s_2^*$ and a buy price quote at $v_t - s_2^*$ when the order book is empty. He monitors only if $s_2^* = s_{I2}$.

(iv) The HFM posts a sell price quote of $v_t + s_1^*$ if the order book is empty on the sell side. He posts a buy price quote of $v_t - s_1^*$ if the book is empty on the buy side. He monitors only if $s_1^* = s_{I1}$.

(v) HFBs submit a market order to trade on news in case there is news.

Informed HFMs earn rents from economies of scope. To understand why, consider the case when an informed HFM posts a two-sided order and in the next interval an LT arrives, but there is no news event. At the next arrival the HFM can only post a one-sided order. No competing HFM can post a half-spread less that s_{I1}, the informed one-sided competitive half-spread. It is optimal for the incumbent HFM to then post s_{I1}, as this quote cannot be undercut. He makes a positive expected profit, since $s_{I1} > s_{I2}$. The incumbent HFM benefits from economies of scope which are unavailable to competitor HFMs. The potential for rents disappears if monitoring is never optimal.

3.5 Steady state equilibrium

Proposition 1 stated that there are only two possible equilibrium levels of the half-spread at each side of the order book, i.e., s_2^* or s_1^*. Consequently, after the order submission stage there are four (2^2) possible states of the order book:

$$(Ask_t - v_t, v_t - Bid_t) \in \left\{ \left(s_2^*, s_2^*\right), \quad \left(s_2^*, s_1^*\right), \quad \left(s_1^*, s_2^*\right), \quad \left(s_1^*, s_1^*\right) \right\}. \tag{10}$$

The equilibrium in the trade economy can thus be described by a Markov chain with the state space defined by the state of the order book. Let the distribution across order book states be denoted by the stochastic row vector x. Then,

$$x^t = x^{t+1}P, \tag{11}$$

where P is the transition matrix:

$$P = \begin{pmatrix} (1-\alpha\delta)(1-\mu\delta)+\alpha\delta & (1-\alpha\delta)\frac{1}{2}\mu\delta & (1-\alpha\delta)\frac{1}{2}\mu\delta & 0 \\ \alpha\delta & (1-\alpha\delta)\left(1-\mu\delta+\frac{1}{2}\mu\delta\right) & 0 & (1-\alpha\delta)\frac{1}{2}\mu\delta \\ \alpha\delta & 0 & (1-\alpha\delta)\left(1-\mu\delta+\frac{1}{2}\mu\delta\right) & (1-\alpha\delta)\frac{1}{2}\mu\delta \\ \alpha\delta & 0 & 0 & 1-\alpha\delta \end{pmatrix}. \tag{12}$$

With probability $\alpha\delta$ there is a news event between two consecutive HFM arrivals. Proposition 1 states that the first arriving HFM will post two-sided price quotes and the book jumps to the spread state $\left(s_2^*, s_2^*\right)$.

With probability $(1-\alpha\delta)$ there is no news event between two consecutive HFM arrivals. With probability $\mu\delta$ an LT arrived an, with equal probability, consumed either the bid or the ask quote. The order book is refilled based on a one-sided competitive half-spread. The incumbent HFM earns rents if monitoring is optimal.

The market always remains in the same state if there is no news event or liquidity trader arrival, i.e., with probability $(1-\alpha\delta)(1-\mu\delta)$. This term appears in all diagonal terms of P.

The steady state probability distribution of the book is given by the left eigenvector of P corresponding to the unit eigenvalue ($\lambda P = \lambda$). The eigenvector λ is given by

$$\lambda = \left(\frac{\alpha(2\alpha+\mu-\alpha\mu\delta)}{\mu^2(1-\alpha\delta)^2}\frac{1}{L}, \quad \frac{\alpha}{\mu(1-\alpha\delta)}\frac{1}{L}, \quad \frac{\alpha}{\mu(1-\alpha\delta)}\frac{1}{L}, \quad \frac{1}{L} \right), \tag{13}$$

where $L = (\frac{\alpha(2\alpha+\mu-\alpha\mu\delta)}{\mu^2(1-\alpha\delta)^2} + \frac{2\alpha}{\mu(1-\alpha\delta)} + 1)$ is a normalization factor to guarantee that the probabilities sum up to one.

The steady state spread s is defined by the scalar product of the steady state probability distribution and the spread in each state of the book:

$$s \equiv \lambda \cdot \left(2s_2^*, s_1^* + s_2^*, s_1^* + s_2^*, 2s_1^*\right). \tag{14}$$

Proposition 2 presents the main result of the model. It relates the equilibrium steady state spread to exchange latency and asset volatility.

Proposition 2. (Steady state spread properties) *The steady state equilibrium spread s*

1. *increases in exchange speed (i.e., it decreases in δ).*
2. *increases in the frequency of news arrival (α).*
3. *increases in the size of news (σ).*

Figure 3 illustrates how the steady state spread increases in exchange speed. The moment speed exceeds a threshold, the adverse selection becomes so large that monitoring becomes optimal for an HFM. The slope of the relationship between the steady state spread and exchange speed decreases as monitoring reduces the marginal adverse selection cost.

[insert Figure 3 here]

The effect of exchange latency can be decomposed into a static *conditional spread effect* and a dynamic *spread distribution effect*. The *conditional spread effect* captures the impact of exchange latency on the spread through increased adverse selection cost, keeping the state probabilities constant. The *spread distribution effect* captures the effect of latency on steady state probabilities:

$$\frac{\partial s}{\partial \delta} = \underbrace{\frac{\partial \lambda}{\partial \delta} \cdot \left(2s_2^*, s_1^* + s_2^*, s_1^* + s_2^*, 2s_1^*\right)}_{\text{Spread distribution effect}} + \underbrace{\lambda \cdot \frac{\partial}{\partial \delta}\left(2s_2^*, s_1^* + s_2^*, s_1^* + s_2^*, 2s_1^*\right)}_{\text{Conditional spread effect}}. \tag{15}$$

Proposition 3. (Dynamic and static latency effects) *If HFMs monitor quotes the spread increases as the exchange speed increases both through a static and a dynamic channel. A higher exchange speed both increases the adverse selection cost (conditional spread) and the probability of rents for HFMs as the state*

distribution shifts. If monitoring is never optimal, the exchange latency does not affect the steady state distribution. The static effect is the only one in that case.

[insert Figure 4 here]

The decomposition of the spread change into the two effects is illustrated in Figure 4. The spread distribution effect is generated by monitoring rents. In low latency environments, the probability of a news event between HFM arrivals ($\alpha\delta$) decreases. Market makers compete through two-sided quotes only after a common value innovation (as outstanding quotes are either wiped out through a market order or have become stale). In equilibrium, two-sided quote competition becomes less likely at low latencies. Therefore, the probability of an incumbent HFM earning rents increases. If monitoring is never optimal, rents are not possible. In this case, a higher exchange speed increases the spread only through a widening conditional spread.

3.6 Exchange latency and quote flickering

The Securities and Exchanges Commission (SEC) defines flickering quotes as "*quotations that change multiple times in a single second*."[7] The model predicts that, all else equal, a faster exchange produces more flickering. Suppose that the model parameter δ is expressed as a fraction of a second. Let a low frequency investor monitor the market each second. The quote flickering intensity is higher if the low frequency investor observes a lower fraction of total market activity.

Definition 2. (Quote flickering) Quote flickering is defined as the expected number of spread size changes in a second.

The (unconditional) probability of a spread change between two HFT arrivals is given by the weighted average of transition probabilities to a different state. The weights are the steady state probabilities.

$$Pr[Spread\ change] = \lambda(\iota - diag(P)), \tag{16}$$

[7] Regulation NMS: SEC Exchange Act Release No. 34-51808 (June 9, 2005)

where ι denotes a unit matrix and $diag(P)$ is the matrix with the same diagonal as P but zero elsewhere.

Therefore,

$$E[Spread\ changes\ per\ second] = \frac{Pr[Spread\ change]}{\delta} = \frac{\lambda(\iota - diag(P))}{\delta}. \tag{17}$$

Proposition 4. (Flickering and exchange latency) *Quote flickering increases in exchange speed (i.e., decreases in δ).*

[insert Figure 5 here]

In fast markets, low frequency observers miss more of the order book activity. In the model, high-frequency traders respond only to news and to liquidity driven market orders. The intensity of these market events does not depend on speed. Flickering is thus more intense in faster markets even in the absence of strategic order submission and cancellations, a channel explored in Baruch and Glosten (2013) and Yueshen (2014). Figure 5 illustrates how quote flickering increase in exchange speed.

4 Empirical results

The NASDAQ-OMX speed upgrade, INET, is used as an instrument to test the predictions of the model. The upgrade was implemented on February 8, 2010, on the equity exchanges of Copenhagen, Helsinki, and Stockholm. The round-trip narrow exchange latency dropped from 2.5 milliseconds to 250 microseconds. We expect that the total latency drop for HFTs was be even larger as NASDAQ-OMX made a colocation available as part of the upgrade. The section first generates the most salient predictions of the model, then discusses the data sample with trader identity, provides summary statistics, and finally tests the predictions.

4.1 Empirical predictions

Perhaps the model's most "idiosyncratic" and surprising prediction is that a lower exchange latency increases the adverse selection cost on price quotes from HFMs ($\frac{\partial s}{\partial \delta} < 0$, see Proposition 2). This prediction will be tested in this section utilizing a NASDAQ-OMX speed upgrade.

A natural second prediction on flickering could not be tested as the data only reveals trader identity on transactions, not on price quotes.

4.2 Data sample

Data is collected from the Thomson Reuters Tick History (TRTH) database. The data consists of trade and quote information for the NASDAQ-OMX exchanges in Copenhagen, Helsinki, and Stockholm. A unique feature of this data is that it reveals trader identity for each side of a transaction.

The data sample consists of all 40 stocks that are included in the OMX Nordic 40 index. It covers a period of six months, three months before and after the exchange speed upgrade: November 8, 2009, to May 8, 2010. In this period 111 traders are active. The start of the sample is chosen so as to not overlap with another major infrastructure change on NASDAQ-OMX, the implementation of centralized clearing on October 19, 2009. Appendix B presents a snapshot of the data. Data is aggregated to the stock-day-trader level.

High-frequency traders. High-frequency traders (HFTs) are identified by following the approach proposed by Kirilenko, Kyle, Samadi, and Tuzun (2011). A trader is labeled HFT if he meets essentially two conditions. First, his daily position change does not exceed 5% of his volume. The trader mean-reverts most of his position within the day. Second, the average distance between a trader's minute-end position and end-of-day position does not exceed 1.5% of his volume. The trader keeps his intraday position close to zero. Further details on the methodology are presented in Appendix C.

The identification procedure identifies five high-frequency traders in the data set: Citadel Securities, Spire Europe, International Algorithmic Trading GmbH, Getco Europe, and Nyenburgh Holding B.V. The empirical tests focus on the adverse selection cost on limit orders by these HFTs, i.e., it is the cost they incur as HFMs.

Various variables are used in the panel regression to test the predictions. Their acronym, definition, and computation is summarized in Appendix A. One variable is key in the tests and deserves to be discussed in the remainder of this subsection.

Adverse selection cost. The adverse selection cost on price quotes is derived from a standard effective spread decomposition model (see, e.g., Bessembinder, 2003; Hendershott, Jones, and Menkveld, 2011). For each trade, define p_t as the transaction price and m_t as the prevailing midpoint at the time of the transaction. A transaction sign dummy q_t takes the value one for buys and minus one for sells (from the perspective of the market-order submitter).

The effective spread (ES) for limit orders is defined as the distance of the transaction price from the prevailing midpoint, expressed as a fraction of that midpoint:

$$ES = q_\tau \frac{p_\tau - m_\tau}{m_\tau}, \tag{18}$$

where τ denotes the time of the transaction.

The average effective spread is decomposed into an adverse selection cost component (AS) and a realized spread (RS) "profit" component. A five-minute "wait" (Δ) is used to compute the market order's long-term price impact:

$$ES = \underbrace{q_\tau \frac{m_{\tau+\Delta} - m_\tau}{m_\tau}}_{AS} + \underbrace{q_\tau \frac{p_\tau - m_{\tau+\Delta}}{m_\tau}}_{RS}. \tag{19}$$

The adverse selection component is a cost as it captures the extent to which a price moves against the price quote submitter upon execution.

4.3 Summary statistics

[insert Table 1 here]

Table 1 reports trade statistics for the months before and after the exchange speed upgrade. The table leads to a couple of observations. First, adverse selection cost for price quotes increased by 51%, from 2.63 basis points to 3.99 basis points. Second, conditioning on HFM price quotes only, the adverse selection cost increased by 574%, from 0.39 to 2.59 basis points. The cost increased by more than ten times as much for

HFMs. Note that in spite of this extraordinary increase, the cost remains lower for HFMs potentially due to superior monitor capacity. Third, effective spread for all price quotes increased by 6.8%, from 4.24 basis points to 4.53 basis points. Conditioning on HFM prices quotes, the increase was three times as large, from 4.48 basis points to 5.44 basis points. Fourth, NASDAQ-OMX index volatility remained largely unchanged as it was 0.99% pre-event and 1.01% post-event. Finally, volume increased by 35%, from 121.46 million stocks per day to 164.02 stocks per day.

[insert Figure 6 here]

Figure 6 illustrates that both the adverse selection cost and the effective spread on HFM limit orders increase around the NASDAQ-OMX speed upgrade. One salient feature of the figure is that both measures seem to jump to a higher level on February 1, 2010. This jump occurs a week before the implementation date and is due to the fact that traders were allowed to test the new system in the week before its official launch.[8]

4.4 Panel regression results

A panel data regression with fixed effects is used to test the model's main predictions[9]:

$$AS_{ijt} = \beta_0 + \beta_1 d_j^{HFM} + d_t^{INET}\left(\beta_2 + \beta_3 d^{HFM}\right) + \theta_i + Controls_{it} + \varepsilon_{ijt}, \tag{20}$$

where i indexes stocks, j indexes traders, and t indexes days. The dependent variable is the stock-trader-day average adverse selection cost on limit orders. The dummy d_t^{HFM} is one when an HFT issued the price quote that was consumed and zero otherwise. The dummy d_t^{INET} takes the value one after NASDAQ-OMX switched to INET and is zero otherwise. The main explanatory variable are the HFM dummy and volatility. They are interacted with the INET dummy to be able to test the predictions for the exchange speed upgrade. The control variables that are included stem from standard spread decomposition models. They include volatility, share turnover, the inverse of price, and market capitalization (see, e.g., Hendershott, Jones, and

[8] As confirmed by a NASDAQ-OMX official.

[9] The availability of trader identity only for the "treated sample" precludes adding a non-treated sample (e.g., German stocks). A full diff-in-diff approach is therefore infeasible.

Menkveld, 2011). The regression further includes stock fixed effects θ_i. Standard errors are computed based on clustering across both stocks and days (Petersen, 2009).

[insert Table 2 here]

Table 2 finds that adverse selection cost on HFM price quotes increases after the exchange speed upgrade. Various model specifications show that adverse selection cost is increased for all price quotes after the upgrade. The following detailed picture arises from the most general model, i.e., model (1). The adverse selection cost increase from agents other than the identified HFMs, is significant and amounts to 1.43 basis points. The increase for HFMs is a significant 47% higher, i.e., 1.43 + 0.68 = 2.11 basis points. The effect is economically large as adverse selection cost is more than five times higher for HFMs after the event, i.e., it changes from 0.39 basis points (see Table 1) to 0.39 + 2.11 = 2.50 basis points, an increase of 541%. The adverse selection cost increase confirms the model's main prediction as stated in Subsection 4.1. The control variables – when significant – have the expected sign. For example, the sign is the same as Hendershott, Jones, and Menkveld (2011, Table III).

The panel regression analysis is repeated with the realized spread as the dependent variable. This complements the adverse selection analysis as the realized spread is the other component of the effective spread. The table shows that the overall realized spread significantly decreases by 1.02 basis points after the speed upgrade. For the HFMs the decrease is less pronounced as for them it declines by only 1.02 – 0.37 = 0.65 basis points. The economic magnitude of this change is modest relative to the adverse selection change, i.e., realized spread drops from 4.48 – 0.39 = 4.09 basis points to 4.09 – 0.65 = 3.44 basis points for HFMs, a decline of 16%. The table further shows that in general the realized spread is higher for HFMs when compared to other quote submitters, 2.04 basis points higher in the most general model.

Finally, the effective spread is used as the dependent variable. The general effective spread increase after the speed upgrade is significant and amounts to 0.41 basis points. For HFMs the increase in effective spread is 0.41 + 1.04 = 1.44 basis points. It is economically large as it implies a 32% increase in the effective spread relative to a pre-upgrade level of 4.48 basis points.

In summary, the spread component analysis shows that the speed upgrade increased the effective spread charged by HFMs. The result is both statistically and economically significant. This increase is driven by a very strong increase in the adverse selection cost of their quotes that is partially offset by a decline in their gross profit, i.e., the realized spread.

5 Model calibration

The model is calibrated in this section to analyze whether it could generate the INET introduction pattern for "reasonable" parameter values. We target the adverse selection cost change for HFMs as this result is at the heart of the model. The strategy is to first pick values for the parameter estimates and calculate the adverse selection cost change that it implies. Then, the elasticity of the cost change to each of the parameters is calculated to get a sense for which parameters matter most for the calibration fit.

The following values were picked for the model parameters:

- **Pre- and post-event latency (δ_{pre} and δ_{post}, respectively).** The NASDAQ-OMX INET upgrade was primarily a change in speed. The latency of the system itself was decreased from 2.5 to 0.25 milliseconds. More important was the introduction of colocation, which enabled clients to avoid latency due to traveling a physical distance to the exchange via cable. A latency reduction due to colocation is estimated to be at around 50 milliseconds.[10]

- **Intensity of liquidity trader arrivals (μ).** A stylized fact that emerged from econometric analysis of high-frequency trade data is that trades are extremely clustered in time. Trading is characterized by bursts of activity and long periods of calm (e.g., Engle and Russell, 1998). Inter-trade durations therefore exhibit strong right skewness. We therefore use the median inter-trade duration for trades in which at least one side is a non-HFT as a basis for the calibration. This median duration is 75 milliseconds. It implies a liquidity trader arrival intensity of 13 per second.

- **Intensity and size of news arrivals (α and σ, respectively).** News is any piece of information that is relevant for predicting fundamental price change, i.e., any information sent through newswire

[10] Latency change due to colocation as estimated by Gainsville Data Center (http://goo.gl/ppRr5n).

services, order book activity, same-industry stock activity, index activity, etc. We set it at five items per second, each with a size of 15 basis points. All else equal, these values match the adverse selection cost for HFMs in the post-event period: 2.59 basis points (see Table 1). Admittedly, they were chosen somewhat arbitrarily but the calibration fit turns out not to depend strongly on these values (analysis below).

- **Monitoring cost (c).** The monitoring cast is set to 0.1 basis points per second. As this cost is to be understood as the shadow cost of directing installed computing power and bandwidth towards parsing information for a particular security, it is hard to find a public source for the level of this cost. We are fortunate to again find that the calibration fit is relatively insensitive to this parameter.

At these parameter values the model-implied adverse selection cost increase due to a latency reduction from 50 to 0.25 milliseconds is 40%.[11] If the pre-event latency is increased to 133 milliseconds then there is a perfect match with the observed 541% increase in adverse selection cost. We believe such pre-event latency value is not "crazy." For example, Budish, Cramton, and Shim (2013) find that median duration for an arbitrage opportunity between the most active Chicago index futures and the most active New York index product is about 100 milliseconds in 2006. This value is large relative to its lower bound set by data transmission at the speed of light between New York and Chicago: 3.81 milliseconds.

A sensitivity analysis reveals that the calibration fit is most sensitive to the liquidity trader arrival rate and pre-event latency. We calculate the elasticity of the relative change of adverse selection cost $(AS_{post} - AS_{pre})/AS_{pre}$ to all of the model parameters. The elasticities are: 1.66 with respect to liquidity trader arrival rate μ, 1.49 with respect to pre-event latency δ_{pre}, -0.17 with repect to the news rate α, -0.005 with respect to news size σ, and 0.005 with repect to monitoring cost c.[12]

[11] The adverse-selection part of the model-implied spread is calculated by subtracting the average rent and monitoring cost from the model-implied spread.

[12] For ease of exposition, the probability of an HFB arriving before an HFM on a news event was set to 1/2. It turns out that freeing up this parameter does little for the calibration fit. Its sensitivity is extremely low: -0.001. The reason is that it scales pre- and post-latency adverse selection cost almost equally.

6 Concluding remarks

This paper finds that a reduction in exchange latency could hurt liquidity. The recursive model reveals that trading becomes more of a zero-sum game between high-frequency traders as (by assumption) they are the only ones whose clock speed can match the speed of the exchange. High-frequency market makers have to set a wider spread to recoup the increased adverse selection cost due to more often meeting high-frequency "bandits." The liquidity trader suffers as he has to pay a higher equilibrium spread. The model establishes this result and shows that it operates on two levels: a state by state "static" effect and a "dynamic" effect through a change in the steady state probability distribution. The analysis further reveals an additional social cost. If the exchange speed passes a threshold then a high-frequency market maker will invest in monitoring technology which is a deadweight cost to the trading community.

The empirical analysis exploits a NASDAQ-OMX speed upgrade to test the model's main prediction. The results show that the adverse selection cost for high-frequency market makers indeed jumps to a higher level after the upgrade. This cost increase is reflected in the HFM effective spread which also increases significantly after the upgrade.

The paper's findings contribute to the public debate on electronic markets and, in particular, the role of speed in the trading process. It adds the insight that a faster market implies more interaction among HFTs, i.e., their market participation increases and, more importantly, transaction cost for "low frequency" investors increases as a result.

Finally, the novel model framework could be used for additional economic analysis of speed. The model's relative appeal is two-fold. First, it can be solved in closed-form. Second, its recursive structure allows for identification of dynamic effects. A type of question that could be analysed with this model is how various policies affect liquidity (e.g., minimum resting time for limit orders, financial transaction tax).

References

Aït-Sahalia, Yacine, and Mehmet Saglam, 2014, High frequency traders: Taking advantage of speed, .

Alizadeh, Sassan, Michael W. Brandt, and Francis X. Diebold, 2002, Range-based estimation of stochastic volatility models, *The Journal of Finance* 57, 1047–1091.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys, 2003, Modeling and forecasting realized volatility, *Econometrica* 71, 579–625.

Angel, James J., Lawrence E. Harris, and Chester S. Spatt, 2013, Equity trading in the 21st century: An update, *Working paper*.

Baron, Matthew, Jonathan Brogaard, and Andrei A. Kirilenko, 2012, The trading profits of high-frequency traders, *Working paper*.

Baruch, Shmuel, and Lawrence R. Glosten, 2013, Fleeting orders, *Columbia Business School Research Paper* 13-43.

Bessembinder, Hendrik, 2003, Issues in assessing trade execution costs, *Journal of Financial Markets* 6, 233–257.

Biais, Bruno, and Thierry Foucault, 2014, High frequency traders and market quality, *Bankers, Markets and Investors* 128, 5–19.

——— , and Sophie Moinas, 2013, Equilibrium fast trading, *HEC Paris Research Paper* 968/2013.

Breckenfelder, Johannes, 2013, Competition between high-frequency traders, and market quality, .

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2014, High frequency trading and price discovery, *Review of Financial Studies (forthcoming)*.

Budish, Eric, Peter Cramton, and John Shim, 2013, The high-frequency trading arms race: Frequent batch auctions as a market design response, *Working paper, University of Chicago*.

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

Foucault, Thierry, Johan Hombert, and Ioanid Roşu, 2013, News trading and speed, *HEC Paris Research Paper* 975/2013.

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

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham, 2014, Toxic arbitrage, .

Foucault, Thierry, Ailsa Roell, and Patrik Sandas, 2003, Market making with costly monitoring: An analysis of the soes controversy, *Review of Financial Studies* 16, 345–384.

Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.

Hagstromer, Björn, and Lars Norden, 2013, The diversity of high-frequency traders, *Journal of Financial Markets* 16, 741–770.

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

Hendershott, Terrence, and Pamela C. Moulton, 2011, Automation, speed, and stock market quality: The nyse's hybrid, *Journal of Financial Markets* 14, 568–604.

Hoffmann, Peter, 2013a, Adverse selection, transaction fees, and multi-market trading, *ECB Working Paper Series* No. 1519, –.

——— , 2013b, A dynamic limit order market with fast and slow traders, *ECB Working Paper Series* No. 1526.

Jones, Charles M., 2013, What do we know about high-frequency trading?, *Columbia Business School Research Paper No. 13-11*.

Jovanovic, Boyan, and Albert J. Menkveld, 2011, Middlemen in limit-order markets, *WFA 2011 paper*.

——— , 2014, Endogenous price dispersion, *Working paper, New York University*.

Kirilenko, Andrei A., Albert (Pete) S. Kyle, Mehrdad Samadi, and Tugkan Tuzun, 2011, The flash crash: The impact of high frequency trading on an electronic market, *Working paper, MIT*.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.

Malinova, Katya, Andreas Park, and Ryan Riordan, 2013, Do retail traders suffer from high frequency traders?, *WFA 2013 paper*.

Martinez, Victor H., and Ioanid Roşu, 2013, High frequency traders, news and volatility, *AFA 2013 San Diego Meetings Paper*.

Menkveld, Albert J., 2013, High frequency trading and the *New Market* makers, *Journal of Financial Markets* 16, 714–740.

——— , 2014, High frequency traders and market structure, *The Financial Review (Special Issue: Computerized and High-Frequency Trading)* 49, 333–344.

Moallemi, Ciamac C., and Mehmet Saglam, 2013, The cost of latency in high-frequency trading, *Operations Research* Articles in Advance, 1–17.

Pagnotta, Emiliano, and Thomas Philippon, 2013, Competing on speed, *Working paper, New York University*.

Patterson, Scott, Jenny Strasburg, and Liam Pleven, 2013, High speed traders exploit loophole, *The Wall Street Journal,* May 1.

Patton, Andrew J., 2011, Volatility forecast comparison using imperfect volatility proxies, *Journal of Econometrics* 160, 246–256.

Petersen, Mitchell A., 2009, Estimating standard errors in finance panel data sets: Comparing approaches, *Review of Financial Studies* 22, 435–480.

Riordan, Ryan, and Andreas Storkenmaier, 2012, Latency, liquidity and price discovery, *Journal of Financial Markets* 15, 416–437.

SEC, 2010, Concept release on equity market structure, Release No. 34-61358; File No. S7-02-10.

Ye, Mao, Chen Yao, and Jiading Gai, 2013, The externalities of high frequency trading, *Working paper*.

Yueshen, Bart Zhou, 2014, Queuing uncertainty, *Working paper*.

Appendix

A Notation summary

Model parameters and their interpretation

Parameter	Definition
v_t	Common value of the risky asset at time t.
δ	Exchange latency: the time elapsed between two HFT arrivals.
α	Probability of a common value change per unit of time.
μ	Probability of an LT arrival per unit of time.
σ	Size of the common and private value innovations.
c	Monitoring cost per unit of time.

Variables used in the empirical analysis.
The subscript i indexes stocks, j indexes traders, and t indexes days.

Variable	Description	Computation
AS_{ijt}	Adverse selection cost	$\frac{m_{\tau+\Delta}-m_\tau}{m_\tau}$ (m_τ is quote midpoint, p_τ is trade price, Δ = 5 min).
RS_{ijt}	Realized spread	$\frac{m_{\tau+\Delta}-p_\tau}{m_\tau}$ (m_τ is quote midpoint, p_τ is trade price, Δ = 5 min).
d_t^{INET}	Post-INET dummy	$d^{INET} = 1$ for days after exchange speed upgrade.
d_j^{HFM}	HFM dummy	$d^{HFM} = 1$ for price quote executions of HFTs.
$Volatility_{it}$	Stock-day volatility	$Volatility_{it} = \frac{1}{2\sqrt{\ln(2)}} \ln\left(\frac{\max_\tau(p_{it\tau})}{\min_\tau(p_{it\tau})}\right)$ where $p_{it\tau}$ denotes transaction prices.[13]
$Turnover_{it}$	Stock-day turnover	$Turnover_{it} = \frac{Volume_{it}}{MarketCap_{it}}$.
$LogMarketCap_{it}$	Market capitalization	Natural logarithm of market capitalization.
$InversePrice_{it}$	Inverse price	Inverse of the closing price.

[13] Alizadeh, Brandt, and Diebold (2002) and Andersen, Bollerslev, Diebold, and Labys (2003) argue that this volatility measure, based on the range between intraday high and low prices, is more robust to microstructure noise than intraday realized volatility. Patton (2011) shows it is an unbiased estimator of the true volatility if the stock price follows a geometric Brownian motion.

B Data sample snapshot

Dates are formatted as dd-mm-yyyy. Time is formatted as hh:mm:ss.μs. Prices are expressed in local currency.

Symbol	Date	Time	Type	Price	Volume	Buyer	Bid	Seller	Ask
CARLa.CO	01-11-2009	09:09:12.582	Quote						380
...									
CARLa.CO	01-11-2009	09:15:49.579	Trade	380	150	ALM		NDA	
...									
CARLa.CO	01-11-2009	09:22:15.529	Quote				378.5		

C HFT identification criteria

Let i index stocks (from one to I), t index days (from one to T), and j index traders (from one to J). Two criteria are used to identify the high-frequency traders. To compute them, we only use the stock-day-trader observations for which the traded volume is at least ten shares.

End of day position. The first criterion states that an HFT mean-reverts most of his position within the day. The average end-of-day net position does not exceed 5% of the daily volume. The dummy $EOD\_position_j$ takes the value one if trader j satisfies the end-of-day position criterion:

$$EOD\_position_j = \begin{cases} 1, & \text{if } \frac{1}{TI}\sum_{i=1}^{I}\sum_{t=1}^{T}\left(\frac{|\sum_\tau Volume_{ijt\tau}q_{ijt\tau}|}{\sum_\tau Volume_{ijt\tau}}\right) \leq 5\% \\ 0, & \text{elsewhere} \end{cases}, \tag{21}$$

where τ denotes the time of the transaction, $Volume_{ijt\tau}$ is the number of shares traded, and the transaction dummy $q_{ijt\tau}$ takes the value one for buys and minus one for sells.

Intraday position. The second criterion states that an HFT keeps his intraday position close to zero. The average of the square root of the sum of squared deviations of the net contract holdings for each minute from the net contract holdings at the end of the day does not exceed 1.5% of the daily volume.

An auxiliary variable is needed to compute this criterion. We partition a trading day into minutes, indexed by τ_m, $m \in \{0, 1, ..., M\}$. Net contract holdings are defined as the net number of contracts bought or sold from the beginning of the day until the end of the minute for which the calculation is made:

$$ContractHolding_{ijt}(\tau_m) = \sum_{\tau=0}^{\tau_m} Volume_{ijt\tau}q_{ijt\tau}, \tag{22}$$

where τ denotes the time of the transaction, $Volume_{ijt\tau}$ is the number of shares traded, and the transaction dummy $q_{ijt\tau}$ takes the value one for buys and minus one for sells.

The dummy $Intraday\_position_j$ takes the value one if trader j satisfies the intraday position criterion:

$$Intraday\_position_j = \begin{cases} 1, & \text{if } \frac{1}{TI}\sum_{i=1}^{I}\sum_{t=1}^{T}\left(\frac{\frac{1}{M}\sqrt{\sum_{m=1}^{M}\left(ContractHolding_{ijt}(\tau_m)-ContractHolding_{ijt}(\tau_M)\right)^2}}{\sum_{\tau} Volume_{ijt\tau}}\right)\leq 1.5\% \\ 0, & \text{else} \end{cases}. \quad (23)$$

The trader j is labeled as an HFT if the end-of-day and intraday positions criteria hold simultaneously. Since the empirical tests focus on the adverse selection cost on limit orders, we are capturing their activity as HFMs. The dummy d_j^{HFM} equals one if both $EOD\_position_j$ and $Intraday\_position_j$ equal one:

$$d_j^{HFM} = EOD\_position_j \times Intraday\_position_j. \quad (24)$$

D Proofs

Lemma 1

Proof. Let $\pi_t(s_t)$ be the HFM expected profit at time t as a function of the half-spread s_t.

The expected profit is conditional on the state of the order book: either empty (no quotes), half full (a quote on one side of the book only), or full (quotes on both sides). If the order book is empty, then the expected profit function is given either by equation (2) if the HFM monitors his quotes, or by equation (4) if the HFM does not monitor his quotes. If the order book is half empty, then the expected profit function is given either by equation (6) if the HFM monitors his quote, or by equation (8) if the HFM does not monitor his quote. If the order book is full, then the HFM cannot post new quotes and his profit is zero.

HFM maximizes total expected profit by choosing for each time t the posted half-spread s_t for which a rival HFM would make zero profit:[14]

$$\max_{s_t}\sum_{t}\pi_t(s_t) \text{ subject to:} \quad (25)$$
$$\pi_t(s_t) = 0,\ \forall t \in \{\delta, 2\delta, ..., k\delta, ...\}.$$

The competitive constraint is binding for all t. Hence, if the HFM monitors his quotes, the competitive half-spread is given at any time t by the solution to either equation (2) or equation (6). If the HFM does not monitor the quotes, the competitive half-spread is given by the solution to equation (4) or equation (8).

The HFM expected profit function only depends on the state of the order book. Thus, the half-spread that sets the HFM expected profit to zero also depends only on the order book state. □

Lemma 2

[14] The argument for defining the competitive half-spread as the one for a rival HFM would make zero profit is discussed in Section 3.4

Proof. We compute the partial derivatives with respect to σ of the competitive half-spread functions s_{I2}, s_{I1}, and s_U:

$$\begin{aligned} \frac{\partial s_{I2}}{\partial \sigma} &= \frac{\alpha\,(2-\delta\mu)}{4\mu+\alpha\,(2-\delta\mu)} > 0, \\ \frac{\partial s_{I1}}{\partial \sigma} &= \frac{\alpha\,(2-\delta\mu)}{4\mu+\alpha\,(2-\delta\mu)} > 0, \text{ and} \\ \frac{\partial s_U}{\partial \sigma} &= \frac{\alpha\,(2-\delta\mu)}{2\mu+\alpha\,(2-\delta)} > 0. \end{aligned} \tag{26}$$

Since all partial derivatives are positive, the competitive half-spread increases with σ, the size of an asset value innovation.

Next, the partial derivatives with respect to μ of the competitive half-spread functions s_{I2}, s_{I1}, or s_U are:

$$\begin{aligned} \frac{\partial s_{I2}}{\partial \mu} &= -\frac{4c\,(4-\delta\mu)+2\alpha\sigma}{(4\mu+\alpha\,(2-\delta\mu))^2} < 0, \\ \frac{\partial s_{I1}}{\partial \mu} &= -\frac{8c\,(4-\delta\mu)+\alpha\sigma}{(4\mu+\alpha\,(2-\delta\mu))^2} < 0, \text{ and} \\ \frac{\partial s_U}{\partial \mu} &= -\frac{4\alpha\sigma}{(2\mu+\alpha\,(2-\delta))^2} < 0. \end{aligned} \tag{27}$$

Since all partial derivatives are negative, the competitive half-spread decreases with μ, the probability of an LT arrival per unit of time.

The partial derivatives with respect to α of the competitive half-spread functions s_{I2}, s_{I1}, and s_U are:

$$\begin{aligned} \frac{\partial s_{I2}}{\partial \alpha} &= -\frac{4\,(2-\delta\mu)\,(c-\mu\sigma)}{(4\mu+\alpha\,(2-\delta\mu))^2}, \\ \frac{\partial s_{I1}}{\partial \alpha} &= -\frac{4\,(2-\delta\mu)\,(2c-\mu\sigma)}{(4\mu+\alpha\,(2-\delta\mu))^2}, \text{ and} \\ \frac{\partial s_U}{\partial \alpha} &= \frac{2\mu\sigma\,(2-\delta\mu)}{(2\mu+\alpha\,(2-\delta))^2} > 0. \end{aligned} \tag{28}$$

The competitive half-spread posted by an uninformed HFM (s_U) increases with α. To show that s_{I2} or s_{I1} also increase with α whenever they are larger than s_U, we establish conditions for $s_{I2} < s_U$ and $s_{I1} < s_U$.

From equations (3) and (5), it follows that $s_{I2} < s_U$ if:

$$s_{I2} - s_U = \frac{\alpha\sigma\,(2-\mu\delta)+4c}{4\mu+\alpha\,(2-\mu\delta)} - \frac{\alpha\sigma\,(2-\mu\delta)}{2\mu+\alpha\,(2-\mu\delta)} = \frac{2\,(4c\mu+\alpha\,(2-\delta\mu)\,(2c-\mu\sigma))}{(4\mu+\alpha\,(2-\delta\mu))\,(2\mu+\alpha\,(2-\delta\mu))} < 0. \tag{29}$$

The denominator of the last expression is always positive. The condition can be rewritten as:

$$8c\mu+2\alpha\,(2-\delta\mu)\,(2c-\mu\sigma) < 0 \Leftrightarrow c < \mu\sigma\frac{\alpha\,(2-\delta\mu)}{2\alpha\,(2-\delta\mu)+4\mu} < \mu\sigma. \tag{30}$$

If $s_{I2} < s_U$, then it follows that $c < \mu\sigma$. From equation (28) and $c < \mu\sigma$ it follows further that $\frac{\partial s_{I2}}{\partial \alpha} > 0$.

From equations (5) and (7), $s_{I1} < s_U$ if:

$$s_{I1} - s_U = \frac{\alpha\sigma(2-\mu\delta)+8c}{4\mu+\alpha(2-\mu\delta)} - \frac{\alpha\sigma(2-\mu\delta)}{2\mu+\alpha(2-\mu\delta)} = \frac{2(8c\mu+\alpha(2-\delta\mu)(4c-\mu\sigma))}{(4\mu+\alpha(2-\delta\mu))(2\mu+\alpha(2-\delta\mu))} < 0. \tag{31}$$

The denominator of the last expression is always positive. The condition can be rewritten as:

$$8c\mu+\alpha(2-\delta\mu)(4c-\mu\sigma) < 0 \Leftrightarrow c < \frac{\alpha\mu\sigma(2-\delta\mu)}{4(2\alpha+2\mu-\alpha\delta\mu)} = \mu\sigma\frac{\alpha(2-\delta\mu)}{4\alpha(2-\delta\mu)+8\mu} < \frac{\mu\sigma}{2}. \tag{32}$$

If $s_{I1} < s_U$, then it follows that $c < \frac{1}{2}\mu\sigma$. From equation (28) and $c < \frac{1}{2}\mu\sigma$, it follows that $\frac{\partial s_{I1}}{\partial \alpha} > 0$.

Next, we compute the partial derivatives with respect to δ of the competitive half-spread functions s_{I2}, s_{I1}, or s_U:

$$\begin{aligned}
\frac{\partial s_{I2}}{\partial \delta} &= -\frac{4\alpha\mu(\mu\sigma-c)}{(4\mu+\alpha(2-\delta\mu))^2}, \\
\frac{\partial s_{I1}}{\partial \delta} &= -\frac{4\alpha\mu(\mu\sigma-2c)}{(4\mu+\alpha(2-\delta\mu))^2}, \text{ and} \\
\frac{\partial s_U}{\partial \delta} &= -\frac{2\mu^2\alpha\sigma}{(2\mu+\alpha(2-\delta))^2} < 0.
\end{aligned} \tag{33}$$

The competitive half-spread posted by a non-monitoring HFM (s_U) decreases with δ (increases with exchange speed). If $s_{I2} < s_U$, then $c < \mu\sigma$ and $\frac{\partial s_{I2}}{\partial \delta} < 0$. Hence, the competitive two-sided half-spread posted by a monitoring HFM also decreases with δ. Similarly, if $s_{I1} < s_U$, then $c < \frac{1}{2}\mu\sigma$ and $\frac{\partial s_{I1}}{\partial \delta} < 0$.

To show that $\min\{s_{I1}, s_{I2}, s_U\} < \sigma$, we compute $s_U - \sigma$:

$$s_U - \sigma = -\frac{2\mu\sigma}{2\mu+\alpha(2-\delta\mu)} < 0. \tag{34}$$

Since $s_U < \sigma$, then $\min\{s_{I1}, s_{I2}, s_U\} < \sigma$ regardless of exchange latency or monitoring costs. □

Lemma 3

Proof. From equations (3) and (7), it follows that:

$$s_{I1} - s_{I2} = \frac{\alpha\sigma(2-\mu\delta)+8c}{4\mu+\alpha(2-\mu\delta)} - \frac{\alpha\sigma(2-\mu\delta)+4c}{4\mu+\alpha(2-\mu\delta)} = \frac{4c}{4\mu+\alpha(2-\mu\delta)} > 0, \tag{35}$$

which is true for any $c > 0$ and $\mu\delta < 1$. For a monitoring HFM, the competitive one-sided half-spread posted is always larger than the competitive two-sided half-spread.

From equation (31) it follows that:

$$s_{I1} < s_U \Longrightarrow c \leq c_1(\delta) \equiv \frac{\alpha\mu\sigma(2-\delta\mu)}{4(2\alpha+2\mu-\alpha\delta\mu)}. \tag{36}$$

The monitoring cost threshold $c_1(\delta)$ decreases with δ.

Let $\underline{c}_1 = \min_\delta c_1(\delta)$, for $\delta = \frac{1}{\mu}$. It follows that $\underline{c}_1 = \frac{\alpha\mu\sigma}{4(\alpha+2\mu)}$. For any $c < \underline{c}_1$, s_{I1} is smaller than s_U regardless of the level of δ. Thus, HFMs monitor one-sided quotes regardless of the exchange speed.

Similarly, let $\overline{c}_1 = \max_\delta c_1(\delta)$, for $\delta = 0$. It follows that $\overline{c}_1 = \frac{\alpha\mu\sigma}{4(\alpha+\mu)}$. For any $c > \overline{c}_1$, s_{I1} is larger than s_U regardless of the level of δ. In this case, HFMs do not monitor one-sided quotes regardless of exchange speed.

For any $c \in \left(\underline{c}_1, \overline{c}_1\right)$, $s_{I1} < s_U$ only if $\delta < \delta_1$, with δ_1 defined below:

$$c - \frac{\alpha\mu\sigma(2-\delta\mu)}{4(2\alpha+2\mu-\alpha\delta\mu)} < 0 \Leftrightarrow \delta < \delta_1 \equiv \frac{1}{\mu}\left[2 - \frac{4c\mu}{\alpha(0.5\mu\sigma - 2c)}\right]. \tag{37}$$

HFMs monitor one-sided quotes only for fast enough markets: $\delta < \delta_1$, which is equivalent to $c < c_1(\delta)$.

A similar reasoning applies for the relationship between the two-sided informed competitive half-spread and the uninformed competitive half-spread. From equation (29) it follows that:

$$s_{I2} < s_U \Longrightarrow c \leq c_2(\delta) \equiv \frac{\alpha\mu\sigma(2-\delta\mu)}{2(2\alpha+2\mu-\alpha\delta\mu)}. \tag{38}$$

Again, the monitoring cost threshold $c_2(\delta)$ decreases with δ.

Let $\underline{c}_2 = \min_\delta c_2(\delta)$, for $\delta = \frac{1}{\mu}$. It follows that $\underline{c}_2 \equiv \frac{\alpha\mu\sigma}{2(\alpha+2\mu)}$. For any $c < \underline{c}_2$, s_{I2} is smaller than s_U regardless of the level of δ. Hence, HFMs monitor two-sided quotes regardless of exchange speed.

Similarly, let $\overline{c}_2 = \max_\delta c_2(\delta)$, for $\delta = 0$. It follows that $\overline{c}_2 = \frac{\alpha\mu\sigma}{2(\alpha+\mu)}$. For any $c > \overline{c}_2$, s_{I2} is larger than s_U regardless of the level of δ. In this case, HFMs do not monitor two-sided quotes regardless of exchange speed.

For any $c \in \left(\underline{c}_2, \overline{c}_2\right)$, $s_{I2} < s_U$ only if $\delta < \delta_2$, with δ_2 defined below:

$$c - \frac{\alpha\mu\sigma(2-\delta\mu)}{4(2\alpha+2\mu-\alpha\delta\mu)} < 0 \Leftrightarrow \delta < \delta_2 \equiv \frac{1}{\mu}\left[2 - \frac{4c\mu}{\alpha(\mu\sigma - 2c)}\right]. \tag{39}$$

HFMs monitor two-sided quotes only for fast enough markets: $\delta < \delta_2$, which is equivalent to $c < c_2(\delta)$.

The monitoring cost threshold values satisfy the inequality $\underline{c}_1 < \overline{c}_1 < \underline{c}_2 < \overline{c}_2$. For $c \in \left(\overline{c}_1, \underline{c}_2\right)$, it follows that $s_{I2} < s_U$ and $s_{I1} > s_U$. It follows that for $c \in \left(\overline{c}_1, \underline{c}_2\right)$, HFMs monitor two-sided quotes but not one-sided quotes. □

Proposition 1

Proof. We prove separately each part of Proposition 1.

Proof for part (i). Let an HFM have an outstanding quote on the news side at the half spread s. If no LT arrives on the news side within the δ interval, then an HFM who does not try to cancel his quote is *always*

adversely selected by HFB. By contrast, an HFM who rushes to cancel a quote on the news side is adversely selected by the HFB only *half* of the time. From equations (2) and (4), it follows that by not rushing to cancel a quote on the news side, the HFM's expected profit drops by:

$$\Delta\pi(s) = \frac{1}{2}\left(1 - \frac{\mu\delta}{2}\right)(\sigma - s) > 0. \tag{40}$$

If the HFM does not cancel a quote on the no-news side, then he will not trade in the next period. Since the common value jumps in the direction of the news, the quote on the no-news side is now at a distance of 2σ from the new common value. The HFM is weakly better off by canceling the no-news side quote and posting a new one at the equilibrium half-spread.

Proof for part (ii). In the absence of a news event, the HFM does not face adverse selection risk from the HFB. Moreover, as the common value of the asset is unchanged, the competitive half-spread values remain the same. If the HFM has either a two-sided quote or a one-sided quote with half-spread s_{I1} outstanding, then he has no incentive to cancel it. In the order submission stage, it is optimal to post the same quote again. If the HFM has a one-sided quote outstanding with a half-spread of s_{I2}, then he can earn a positive profit in the next trading game by filling the other side of the book at the half-spread $s_{I1} > s_{I2}$. No competitor HFM can undercut the s_{I1} half-spread and break even with a one-sided quote.

Proof for parts (iii) and (iv). The size of the half-spread is given by the solution to the zero-profit conditions, explicitly stated in equations (3), (5), and (7). The HFM always posts the minimum competitive half-spread across monitoring strategies. Any other half-spread can be undercut by a competitor HFM who chooses the monitoring strategy that yields the minimum competitive half-spread.

The HFM quotes always add the competitive half-spreads to the current common value. Assume an HFM posts a different quote, with a slightly higher half-spread on the ask side of the book ($\epsilon > 0$):

$$(Ask_t - v_t, v_t - Bid_t) = \{s + \epsilon, s\}.$$

This HFM strategy is suboptimal. A competitor HFM can post a quote at $(Ask_t - v_t, v_t - Bid_t) = \{s, s\}$, and thus undercut the ask quote of the incumbent HFM. The reasoning is identical for the bid side of the book.
Proof for part (v). If there is news, the expected profit of HFB from submitting a market order is:

$$\pi_{HFB}(s) = \frac{1}{2}(\sigma - s) > 0, \tag{41}$$

since Lemma 2 implies that $\min\{s_{I1}, s_{I2}, s_U\} < \sigma$.

□

Proposition 2

Proof. There are three possible values of the steady state spread, depending on the relationship between the competitive half-spreads s_{I2}, s_{I1}, and s_U.

Case 1. If $s_{I2} < s_{I1} < s_U$, the conditional half-spread is the competitive half-spread that corresponds to the case where HFM monitors all quotes: $s_2^* = s_{I2}$ and $s_1^* = s_{I1}$. The equilibrium steady state spread s is given by:

$$s = 2\frac{8c\mu + \alpha^2(2-\delta\mu)^2\sigma + \alpha(8c(1-\delta\mu) + \mu(2-\delta\mu)\sigma)}{(4\mu + \alpha(2-\delta\mu))(\mu + \alpha(2-\delta\mu))}. \tag{42}$$

The partial derivative of the steady state spread with respect to δ is:

$$\frac{\partial s}{\partial \delta} = -2\frac{\mu(\mu\sigma - 2c) + 2\alpha(c + 2c\delta\mu + \mu(2-\delta\mu)\sigma) + \alpha^2(2-\delta\mu)(2c\mu + (2-\delta\mu)\sigma)}{(4\mu + \alpha(2-\delta\mu))^2(\mu + \alpha(2-\delta\mu))^2}. \tag{43}$$

From equation (29), $s_{I2} < s_U$ implies $\mu\sigma - 2c > 0$. Thus, it follows that $\frac{\partial s}{\partial \delta} < 0$. The steady state spread decreases with the exchange latency (increases with exchange speed).

The partial derivative of the steady state spread with respect to σ is positive – the steady state spread increases with the size of asset value innovations:

$$\frac{\partial s}{\partial \sigma} = \frac{2\alpha(2-\delta\mu)}{4\mu + \alpha(2-\delta\mu)} > 0. \tag{44}$$

The partial derivative of the steady state spread with respect with α is given by:

$$\frac{\partial s}{\partial \alpha} = -\frac{(2-\delta\mu)^2\left(2\alpha(2c-\mu\sigma) + \alpha^2((2c-\sigma)(1-\delta\mu) - \mu\sigma(\delta\mu - 3))\right) - \mu^2(2c(6-\delta\mu) + \mu\sigma(2-\delta\mu))}{2(4\mu + \alpha(2-\delta\mu))^2(\mu + \alpha(2-\delta\mu))^2}. \tag{45}$$

Since $s_{I2} < s_U$ implies $\mu\sigma - 2c > 0$, all terms in the numerator are negative. It follows that s is increasing in α.

Case 2. If $s_{I2} < s_U < s_{I1}$, the conditional half-spread is the competitive half-spread that corresponds to the case where HFM monitors only two-sided quotes: $s_2^* = s_{I2}$ and $s_1^* = s_U$. The equilibrium steady state spread s is given by:

$$s = 2\frac{\alpha\left(-\alpha^2(2-\delta\mu)^3\sigma + \alpha(2-\delta\mu)(8c + \mu(6-5\delta\mu)\sigma) + 4\mu(4c + \mu(2-\delta\mu))\sigma\right)}{(4\mu + \alpha(2-\delta\mu))(\mu + \alpha(2-\delta\mu))(2\mu + \alpha(2-\delta\mu))}. \tag{46}$$

To prove s decreases with δ, we first state a helpful result. The second derivative of the steady state spread with respect to the exchange latency and monitoring cost is given by:

$$\frac{\partial^2 s}{\partial\delta\partial c} = \frac{16\alpha^2\mu(5\mu + \alpha(4-2\delta\mu))}{(4\mu + \alpha(2-\delta\mu))^2(\mu + \alpha(2-\delta\mu))^2} > 0. \tag{47}$$

From the proof of Lemma 3, $\bar{c}_2 = \frac{\alpha\mu\sigma(2-\delta\mu)}{2(2\alpha+2\mu-\alpha\delta\mu)}$ is the maximum monitoring cost level for which HFM monitors two-sided quotes. It remains to show that $\frac{\partial s}{\partial \delta} < 0$ for $c = \bar{c}_2$. Then, from equation (47), the steady state spread decreases with latency for any lower monitoring cost.

Indeed,

$$\frac{\partial s}{\partial \delta}(c = \bar{c}_2) = -\frac{4\alpha\mu^2\sigma}{(2\mu + \alpha(2-\delta\mu))^2} < 0. \tag{48}$$

It follows that the steady state spread decreases with latency (increases with the exchange speed) for any monitoring cost for which $s_{I2} < s_U < s_{I1}$.

The steady state spread increases with σ:

$$\frac{\partial s}{\partial \sigma} = \frac{2\alpha(2-\delta\mu)\left(4\mu^2 + \alpha\mu(6-5\mu) + \alpha^2(2-\delta\mu)^2\right)}{(4\mu+\alpha(2-\delta\mu))(2\mu+\alpha(2-\delta\mu))(\mu+\alpha(2-\delta\mu))} > 0.$$

The partial derivative with respect to α is positive for any monitoring cost $c < \bar{c}_2$. Again, we compute the second derivative of the steady state spread with respect to α and c:

$$\frac{\partial^2 s}{\partial\alpha\partial c} = -\frac{16\left(4\mu^2 + \alpha^2(2-\delta\mu)^2\right)}{(\mu+\alpha(2-\delta\mu))^2(4\mu+\alpha(2-\delta\mu))^2} < 0. \tag{49}$$

To show that $\frac{\partial s}{\partial \alpha} > 0$ for any $c < \bar{c}_2$, it remains to verify that $\frac{\partial s}{\partial \alpha} > 0$ for $c = \bar{c}_2$. Indeed,

$$\frac{\partial s}{\partial \alpha}(c = \bar{c}_2) = -\frac{4\mu\sigma(2-\delta\mu)}{(2\mu+\alpha(2-\delta\mu))^2} > 0. \tag{50}$$

Hence, the steady state spread increases with the probability of an asset value innovation for any monitoring cost for which $s_{I2} < s_U < s_{I1}$.

Case 3. If $s_U < s_{I2} < s_{I1}$, the conditional half-spread is the competitive half-spread that corresponds to the case where HFM never monitors. The equilibrium steady state spread is then $2s_U$, as all conditional half-spreads are equal to s_U. Lemma 2 establishes that s_U is increasing in exchange latency, news size, and the probability of a news event. □

Proposition 3

Proof. The effect of exchange latency on the steady state spread is decomposed into a static *conditional spread effect* and a dynamic *spread distribution effect*:

$$\frac{\partial s}{\partial \delta} = \underbrace{\frac{\partial \lambda}{\partial \delta}\cdot\left(2s_2^*, s_1^*+s_2^*, s_1^*+s_2^*, 2s_1^*\right)}_{\text{Spread distribution effect}} + \underbrace{\lambda\cdot\frac{\partial}{\partial\delta}\left(2s_2^*, s_1^*+s_2^*, s_1^*+s^*2, 2s_1^*\right)}_{\text{Conditional spread effect}}. \tag{51}$$

From Lemma 2, all components of the conditional spread vector $\left(2s_2^*, s_1^*+s^*2, s_1^*+s_2^*, 2s_1^*\right)$ decrease with δ. It follows that the static effect is negative. Keeping the state probabilities constant, a lower latency leads to a higher spread.

To assess the sign of the dynamic effect, we compute $\frac{\partial\lambda}{\partial\delta}\cdot\left(2s_2^*, s_1^*+s_2^*, s_1^*+s_2^*, 2s_1^*\right)$ for all possible relationships between the competitive half-spreads s_{I2}, s_{I1}, and s_U.

Case 1. If $s_{I2} < s_{I1} < s_U$, the conditional half-spread is the competitive half-spread that corresponds to the case where HFM monitors: $s_2^* = s_{I2}$ and $s_1^* = s_{I1}$. It follows that:

$$\frac{\partial\lambda}{\partial\delta}\cdot\left(2s_2^*, s_1^*+s_2^*, s_1^*+s_2^*, 2s_1^*\right) = -\frac{16\alpha^2 c\mu}{(\mu+\alpha(2-\delta\mu))^2(4\mu+\alpha(2-\delta\mu))} < 0. \tag{52}$$

The dynamic effect of latency on the steady state spread is also negative. A higher exchange speed leads to a wider spread.

Case 2. If $s_{I2} < s_U < s_{I1}$, the conditional half-spread is the competitive half-spread that corresponds to the case where HFM monitors only two-sided quotes: $s_2^* = s_{I2}$ and $s_1^* = s_U$. It follows that:

$$\frac{\partial \lambda}{\partial \delta} \cdot \left(2s_2^*, s_1^* + s_2^*, s_1^* + s_2^*, 2s_1^*\right) = \frac{8\alpha^2\mu\left(4c\mu + \alpha\left(2-\delta\mu\right)\left(2c-\mu\sigma\right)\right)}{\left(4\mu+\alpha\left(2-\delta\mu\right)\right)\left(2\mu+\alpha\left(2-\delta\mu\right)\right)\left(\mu+\alpha\left(2-\delta\mu\right)\right)^2}. \tag{53}$$

For $s_{I2} < s_U$ to be true, the proof of Lemma 3 shows that the monitoring cost level needs to be low enough: $c < \frac{\alpha\mu\sigma(2-\delta\mu)}{2(2\alpha+2\mu-\alpha\delta\mu)}$. It follows that $\frac{\partial \lambda}{\partial \delta} \cdot \left(2s_2^*, s_1^* + s_2^*, s_1^* + s_2^*, 2s_1^*\right) < 0$. Again, a higher exchange speed leads to a wider spread.

Case 3. If $s_U < s_{I2} < s_{I1}$, the conditional half-spread is the competitive half-spread that corresponds to the case where HFM never monitors: $s_2^* = s_U$ and $s_1^* = s_U$. Then:

$$\frac{\partial \lambda}{\partial \delta} \cdot \left(2s_2^*, s_1^* + s^*2, s_1^* + s^*2, 2s_1^*\right) = 0, \tag{54}$$

since the state probabilities add up to one by definition. Since the spread is the same for all states of the order book, it follows that the sum of the elements of $\frac{\partial \lambda}{\partial \delta}$ is zero. In this case, there is no dynamic effect of latency on the steady state spread. □

Proposition 4

Proof. We compute the partial derivatives with respect to δ for the spread and the quote flickering measures. For the quote flickering measure:

$$\frac{\partial E[\textit{Spread changes per second}]}{\partial \delta} = -\frac{\alpha^3\mu\left(5\mu^2 + 2\alpha\mu\left(6-5\delta\mu\right) + \alpha^2\left(8-12\delta\mu+5\delta^2\mu^2\right)\right)}{\left(\alpha+\mu-\alpha\delta\mu\right)^2\left(\mu+8\left(2-\delta\mu\right)\right)^2} < 0, \tag{55}$$

since the polynomial $8 - 12\delta\mu + 5\delta^2\mu^2$ is larger than zero for any $\delta \in (0,1)$, $\mu \in (0,1)$, and $\delta\mu \leq 1$.

Since *E[Spread changes per second]* decreases with exchange latency δ, flickering increases with exchange speed. □

Table 1: **Summary statistics**

This table presents the mean and its standard deviation for adverse selection cost for price quotes, the effective spread, market index volatility, and volume. Adverse selection is calculated as how much the midquote changes against the price quote after execution (i.e., midquote increase after ask quote execution and midquote decrease after bid quote execution. The waiting period is set to five minutes. The statistics are presented for a sample of OMX Nordic index stocks. The pre-event period runs from November 8, 2009 to February 1, 2010; the post-event period runs from February 8, 2010 to May 8, 2010.

Variable	Sample	Before INET	After INET
Adverse selection (basis points)	All traders	2.63 (0.36)	3.99 (1.16)
	HFM	0.39 (1.11)	2.59 (1.07)
Effective spread (basis points)	All traders	4.24 (0.18)	4.53 (0.39)
	HFM	4.48 (0.93)	5.44 (0.74)
OMX Nordic 40 index daily volatility (percent)	Full sample	0.99 (0.38)	1.01 (0.50)
Daily volume (million stocks)	Full Sample	121.46 (45.31)	164.02 (63.09)

Table 2: **Spread components change after exchange speed upgrade (INET)**

This table presents the results of regressions that relate the effective spread, the adverse selection cost, and the realized spread for price quotes to various explanatory variables. The effective spread is the relative distance between the transaction price and the midquote prevailing at the time of the transaction. Adverse selection is calculated as how much the midquote changes against the price quote after execution (i.e., midquote increase after ask quote execution and midquote decrease after bid quote execution. The waiting period is set to five minutes. ions that relates the realized spread for price quotes to various explanatory variables. The realized spread is the effective spread minus the adverse selection cost (and could be interpreted as a gross profit for the price quote submitter). The explanatory variables are a dummy for the post-event period (d^{INET}), a dummy for high-frequency market makers (d^{HFM}), volatility ($Volatility_{it}$), share turnover ($Turnover_{it}$), log market cap ($LogMarketCap_{it}$), the inverse of the stock price ($InversePrice_{it}$), as well as stock fixed effects. The pre-event period runs from November 8, 2009 to February 1, 2010; the post-event period runs from February 8, 2010 to May 8, 2010. All variables are standardized to have zero mean and unit variance. Standard errors are double-clustered at day and stock levels following Petersen (2009). Robust t-statistics are reported below the coefficients (significance levels are as follows: 1% - ***, 5% - **, 10% - *).

	Adverse selection		Realized spread		Effective spread	
Controls	Yes	No	Yes	No	Yes	No
$d_{HFM}d_{INET}$	0.68^{**} 2.20	0.68^{**} 2.22	0.37 1.22	0.38 1.21	1.04^{***} 4.87	1.06^{***} 4.46
d_{INET}	1.43^{***} 12.06	1.21^{***} 9.42	-1.02^{***} −7.14	-1.12^{***} −6.14	0.41^{**} 2.26	0.08 0.35
d_{HFM}	-1.88^{***} −7.48	-1.89^{***} −7.49	2.04^{***} 7.93	2.03^{***} 7.69	0.15 0.94	0.13 0.75
$Volatility_{it}$	0.34^{***} 7.21		0.03 0.73		0.38^{***} 6.67	
$Turnover_{it}$	−0.06 −1.10		0.04 0.63		−0.02 −0.27	
$LogMarketCap_{it}$	-3.44^{*} −1.79		−4.55 −1.29		−7.98 −1.52	
$InversePrice_{it}$	−1.73 −1.13		−2.55 −1.18		−3.79 −1.19	
Intercept	2.78^{***} 24.11	2.85^{***} 21.69	2.56^{***} 14.09	2.54^{***} 13.26	5.34^{***} 20.89	5.40^{***} 19.33
Stock fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
#Observations	151,075	151,075	151,075	151,075	151,075	151,075

Figure 1: **Optimal monitoring strategy as a function of exchange latency and monitoring cost** This figure plots the optimal monitoring strategy in a two-dimensional space defined by monitoring cost (horizontal axis) and exchange latency (inverted vertical axis). High-frequency market makers (HFMs) always monitor their quotes for low monitoring cost. For medium monitoring cost, monitoring is optimal only if exchange latency is low enough. For high monitoring cost HFMs never monitor, irrespective of exchange latency.

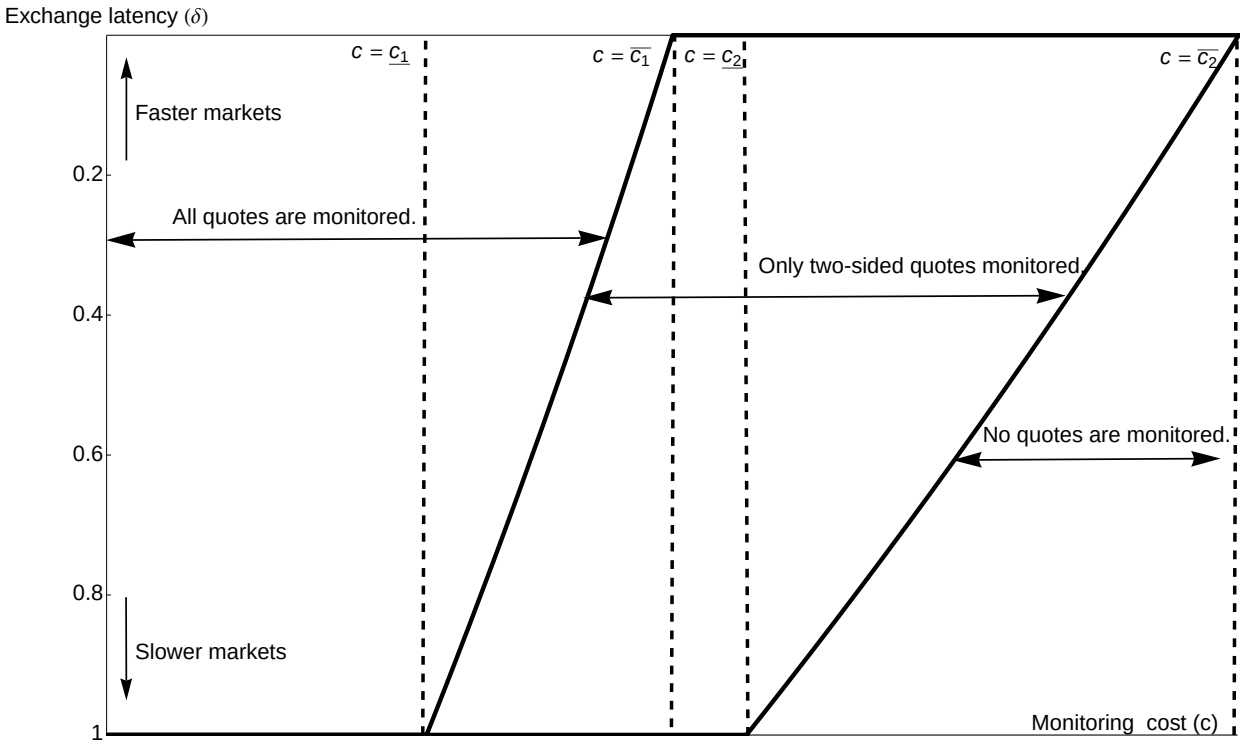
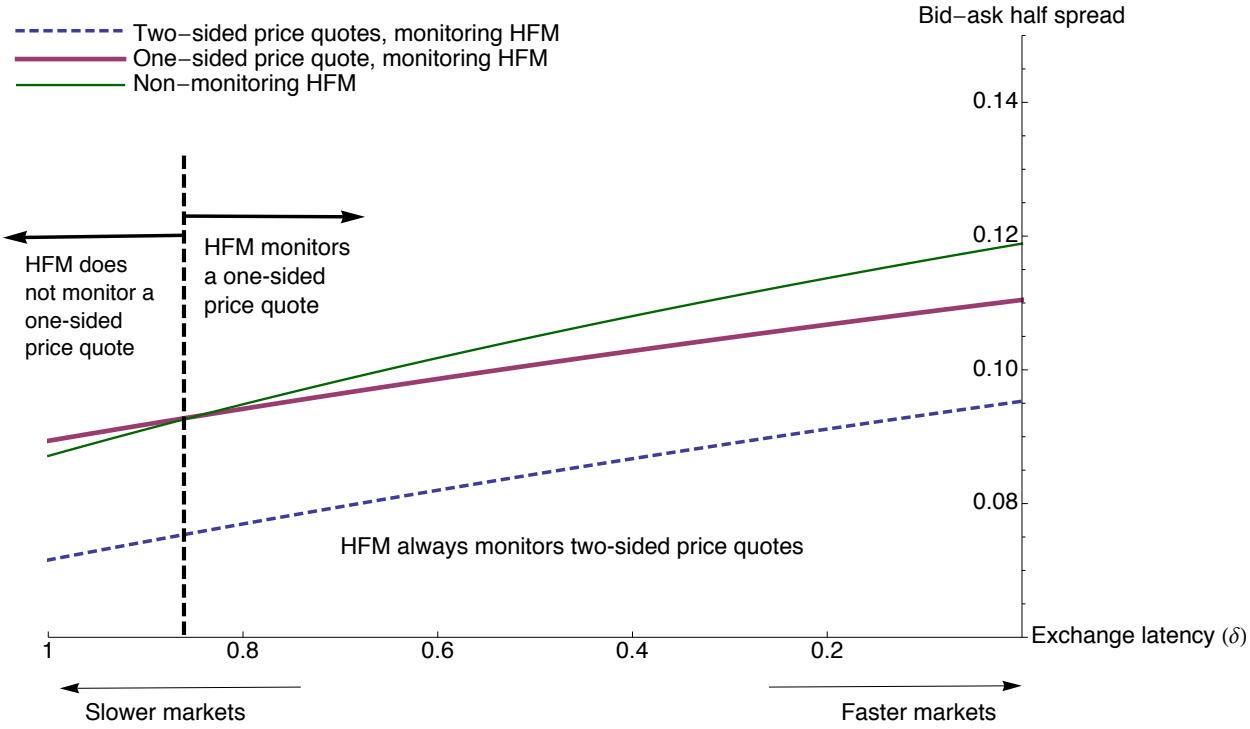
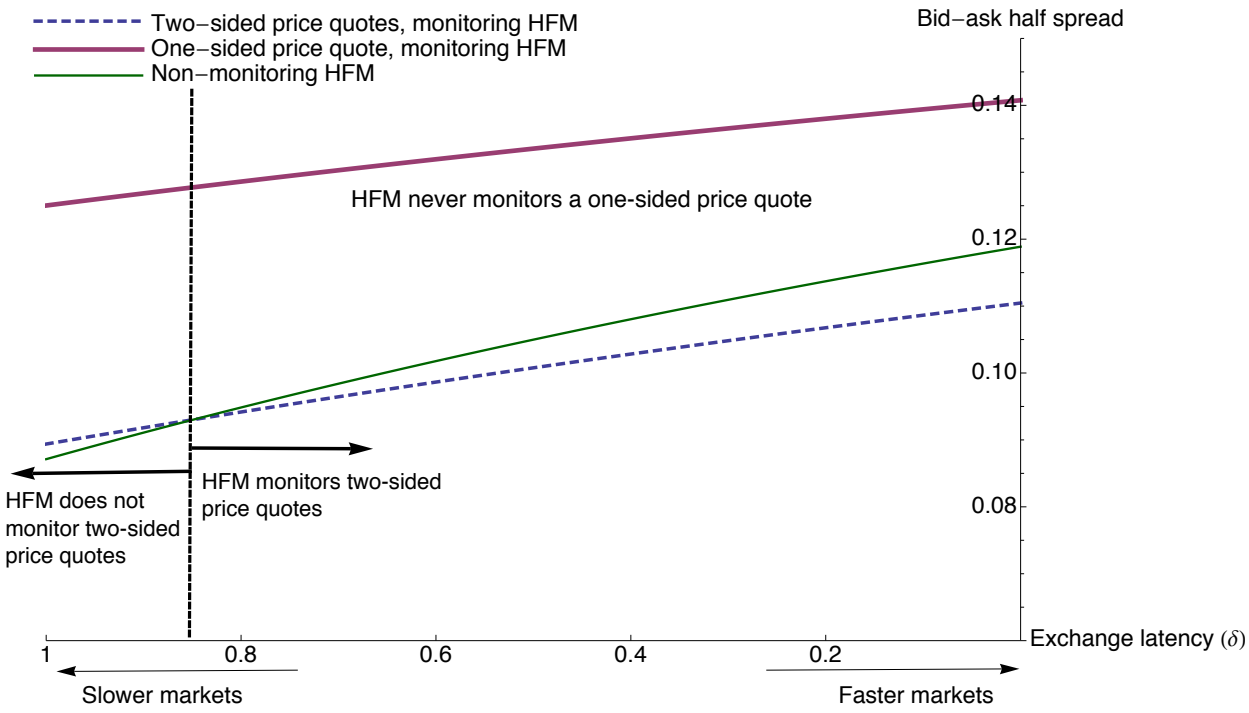


Figure 2: **Competitive bid-ask spread and exchange latency**

This figure illustrates the optimal monitoring strategy for high-frequency market makers (HFMs). The strategy is pinned down by competitive pressure of rival HFMs. It therefore determined by the lower envelope of the half-spread curves associated with a monitoring and non-monitoring HFM. Panel (a) shows that for low monitoring cost HFMs decide to always monitor their two-sided price quotes if they find an empty book on arrival. If they find a one-sided book they decide to monitor only if the exchange is fast enough. Panel (b) finds that for high monitoring cost HFMs monitor two-sided price quotes only if the exchange is fast enough. They never monitor a one-sided price quote. Parameter values are: $\alpha = 0.92$, $\mu = 0.86$, and $\sigma = 0.23$.



(a) Low monitoring cost ($c = 0.025$)



(b) High monitoring cost ($c = 0.05$)

Figure 3: **Steady state bid-ask spread and exchange latency**

This figure depicts how the equilibrium bid-ask spread changes with exchange latency for low and high monitoring cost. The spread is calculated as a weighted average where the weights are the steady state probabilities. The vertical lines correspond to the exchange latency levels for which the optimal HFM monitoring strategy changes. Dotted lines correspond to a high monitoring cost and solid lines to a low monitoring cost. Parameter values are: $\alpha = 0.92$, $\mu = 0.86$, and $\sigma = 0.23$.

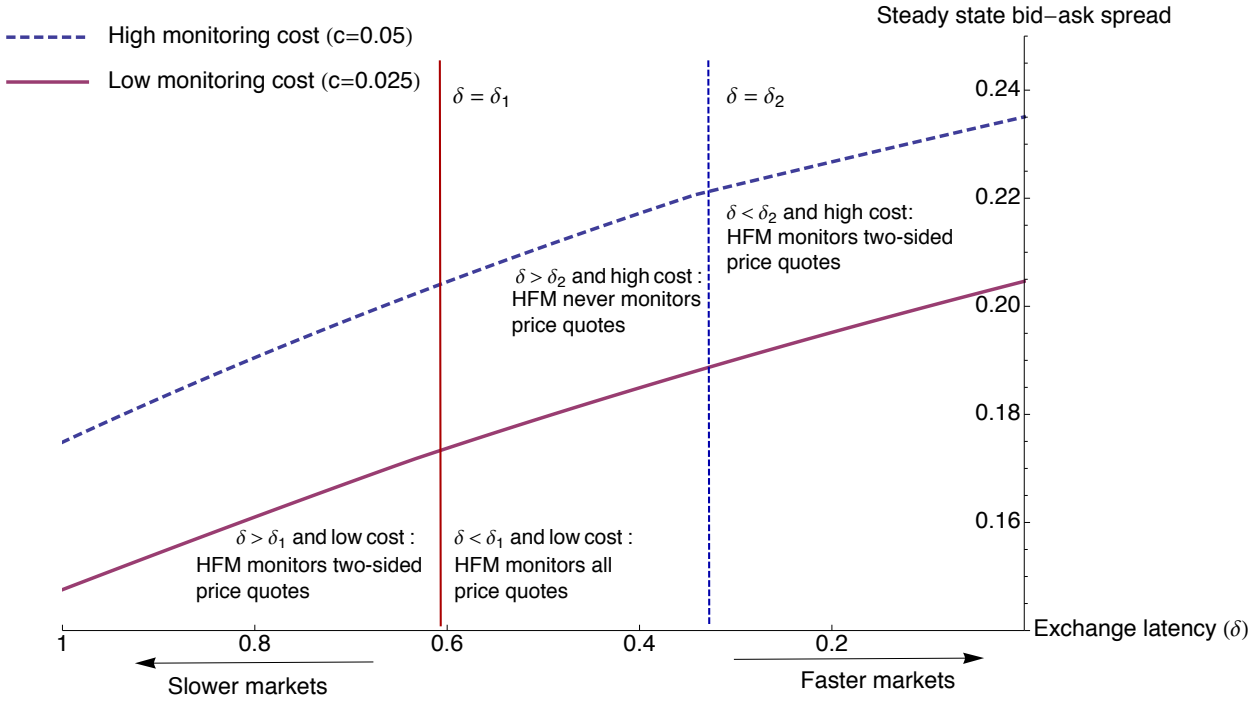


Figure 4: **Decomposition of the effect of exchange latency on the steady state bid-ask spread**
This figure illustrates the decomposition of the effect of exchange latency on the steady state bid-ask spread into a static and a dynamic component. The static component is due to a change in the quoted spread in each state of the world. It is illustrated by keeping the steady state probabilities fixed at the values one observes for the slowest possible exchange ($\delta = 1$). The dynamic component is due to changes in the steady state probabilities as a result of a faster exchange. The size of this additional effect is illustrated by the difference between the steady state spread line and the spread line with fixed steady state probabilities. Parameter values are: $\alpha = 0.92$, $\mu = 0.86$, $\sigma = 0.23$, and $c = 0.05$.

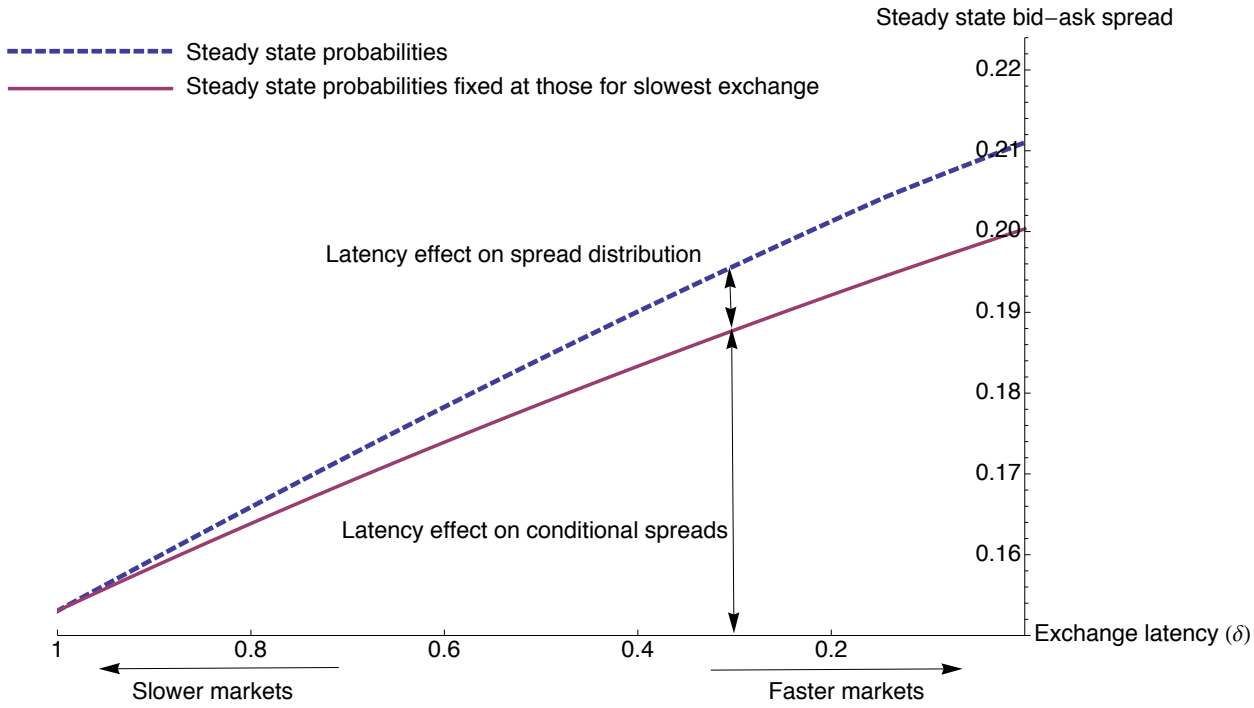


Figure 5: **Flickering as a function of exchange latency**

This figure illustrates how quote flickering depends on exchange latency. Quote flickering is defined as the expected number of spread changes per unit of time. Parameter values are: $\alpha = 0.92$, $\mu = 0.86$, $\sigma = 0.23$, and $c = 0.05$.

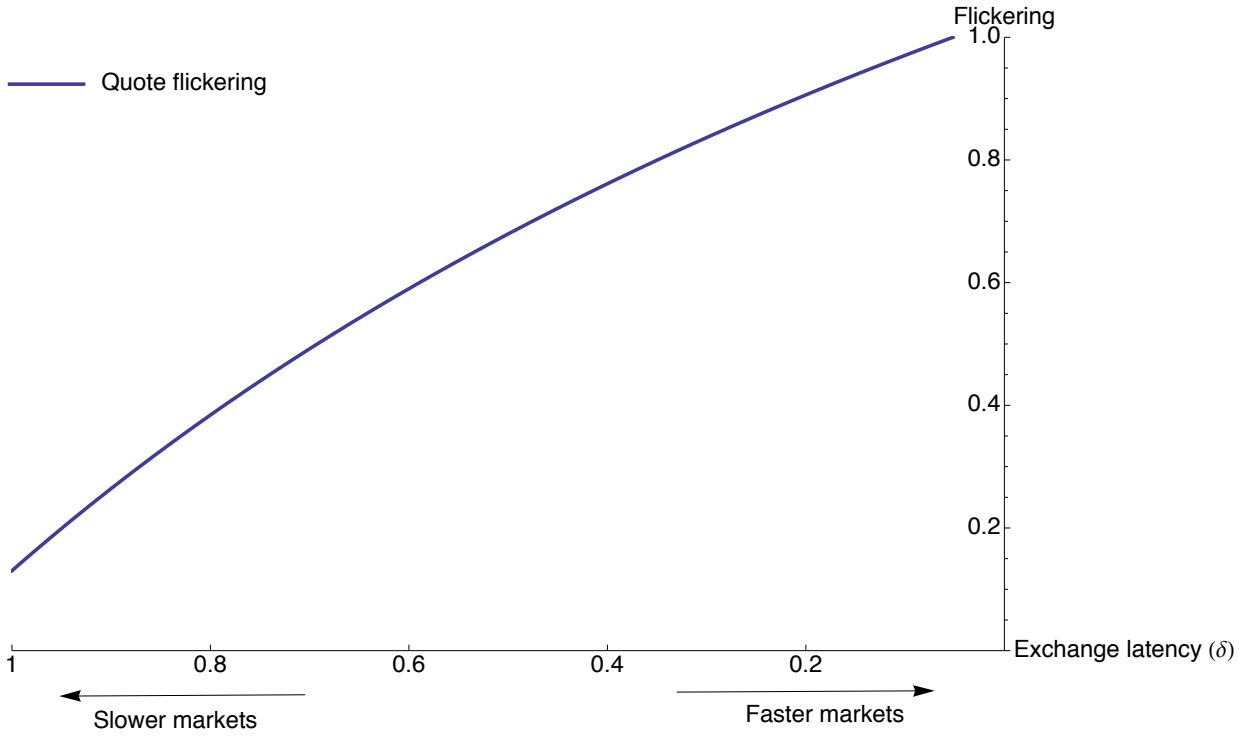
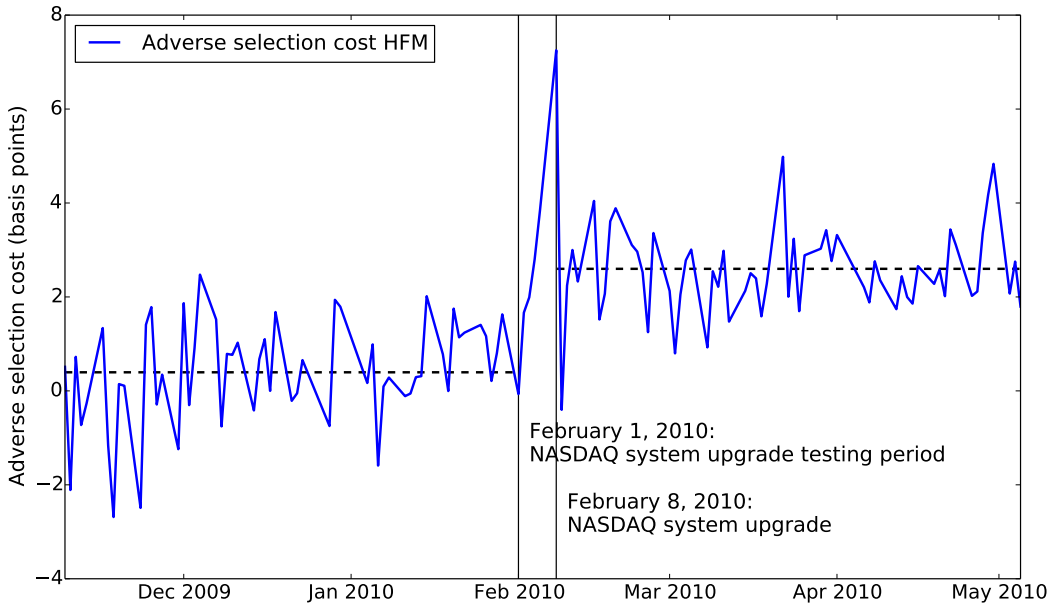
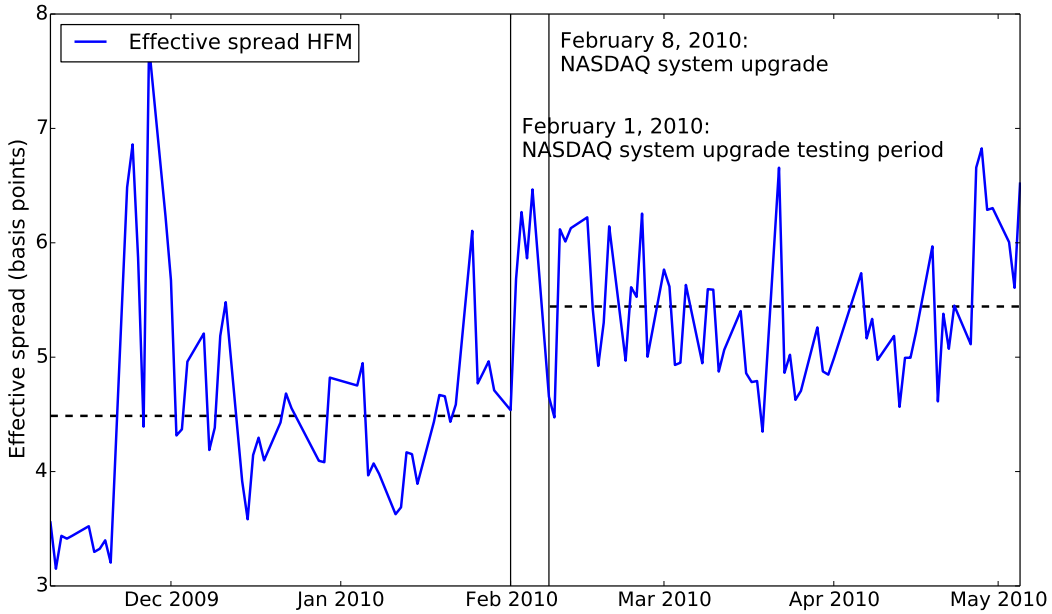


Figure 6: **Adverse selection cost and effective spread before and after exchange speed upgrade** This figure depicts the daily average adverse selection cost and effective spread on price quotes of high-frequency market makers (HFMs) surrounding the NASDAQ-OMX speed upgrade (INET) on February 8, 2010. The adverse selection cost and the effective spread are averaged across all stocks included in the OMX Nordic 40 index. Adverse selection cost is computed as the long-term (five-minute) adverse price impact suffered by price quotes that get executed. The effective spread is the relative distance between the transaction price and the midquote prevailing at the time of the transaction. Traders were allowed to submit orders to the new system in a testing period in the week before the upgrade.



(a) Adverse selection cost



(b) Effective spread