# The Economics of High-Frequency Trading: Taking Stock

Albert J. Menkveld[1,2]

[1]Department of Finance, Vrije Universiteit Amsterdam, Amsterdam 1081 HV, Netherlands

[2]Tinbergen Institute Amsterdam, Amsterdam 1082 MS, Netherlands

Annu. Rev. Financ. Econ. 2016. 8:1–24

First published online as a Review in Advance on September 8, 2016

The *Annual Review of Financial Economics* is online at financial.annualreviews.org

This article's doi:
10.1146/annurev-financial-121415-033010



JEL code: G10

## Keywords

high-frequency trading, electronic markets, microstructure

## Abstract

I review the recent high-frequency trader (HFT) literature to single out the economic channels by which HFTs affect market quality. I first group the various theoretical studies according to common denominators and discuss the economic costs and benefits they identify. For each group, I then review the empirical literature that speaks to either the models' assumptions or their predictions. This enables me to come to a data-weighted judgement on the economic value of HFTs.

## 1. INTRODUCTION

HFT: high-frequency trader

Almost half a century ago, Fischer Black (1971a,b) shared a vision of a fully automated stock exchange, illustrated in **Figure 1**: "It seems likely, then, that . . . a stock exchange can be embodied in a network of computers, and the costs of trading can be sharply reduced, without introducing any additional instability in stock prices, and without being unfair either to small investors or to large investors" (Black 1971b, p. 87). We have come pretty close. Trading floors have been largely replaced by exchange servers. These new venues include centralized limit-order markets, but also crossing networks, dark pools, etc. The services that brokers, dealers, and specialists formerly provided are now largely coded into computer algorithms operating at superhuman speed. Securities trading in all asset classes either has migrated completely to electronic markets (e.g., US equities) or is in the process of migrating (Johnson 2010, Cardella et al. 2014).

One type of computerized trader has attracted the most attention: the high-frequency trader (HFT). A formal definition does not exist, but most observers associate the term HFT with extremely fast computers running algorithms coded by traders who trade for their own account. Collectively, their participation rate in trades is typically a couple of deciles (SEC 2010). These traders typically do not work at deep-pocket sell-side banks, but at privately held firms. They therefore need to keep their positions small and short-lived to keep the capital tied up in margin accounts in check. They trade frequently during the day and avoid carrying positions overnight.


**Figure 1**

Fischer Black's vision of a fully automated exchange appeared in the *Financial Analysts Journal* in 1971 (Black 1971a,b). Figure reproduced with permission.

These characterizations suggest that HFTs are best thought of as a new type of intermediary. Whether HFTs benefit market quality or hurt it is an important question that is fiercely debated in industry, among regulators, in the media, and in a rapidly growing academic literature that I survey in this article.

Before delving into the literature, I would like to share an observation that is often overlooked in the heated debate. HFTs and new venues have helped us migrate quickly to electronic trading, which, in turn, has yielded lower transaction costs and more volume. Let me develop this argument in the next few paragraphs.

Electronic trading, new venues, and HFTs are intimately related. There is arguably a symbiotic relationship between new electronic venues and HFTs. These new venues need HFTs to insert aggressively priced bid and ask quotes, and HFTs need the new venues to satisfy their requirements in terms of automation, speed, and low fees (this is discussed more in Section 3.4).

The electronic market structure that has arisen could be viewed as the product of automation. Replacing humans by machines has yielded cheaper production in many industries, so why not in the industry of securities trading? There is much more to say here about the nature of production in such an industry, but it is useful to at least consider one measurable characteristic of securities markets: end-user transaction cost. Did it indeed decline and, if so, by how much?

**Figure 2** depicts how transaction cost changed for US equities from 2001 to 2011. Arguably the most dramatic market structure changes for US equities all happened in this time span. For

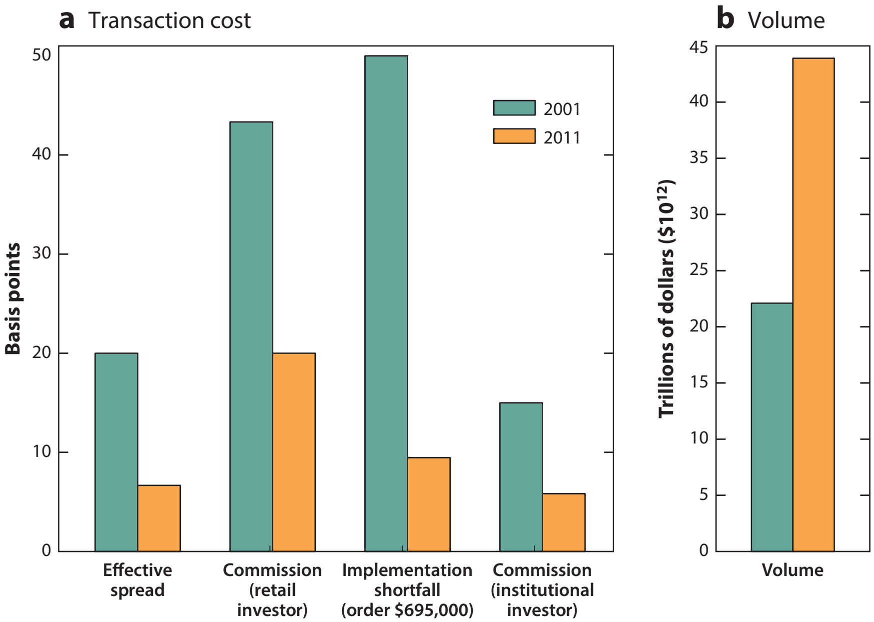


**Figure 2**

This figure plots estimates of implicit and explicit transaction costs for US equity in 2001 and 2011, arguably before and after the entry of high-frequency traders. (*a*) The NYSE/NASDAQ effective spread and retail investor commission were taken from Angel, Harris & Spatt (2015), implementation shortfall from Frazzini (2012), and institutional investor commission from Cappon (2014). Average stock price is assumed to be $30 (Angel, Harris & Spatt 2015, p. 23), and average retail order size is assumed to be $6,000 (Lee & Radhakrishna 2000, p. 102). (*b*) Volume is expressed in February 2016 dollars and was taken from **data.worldbank.org**.

example, the New York Stock Exchange (NYSE) lost its dominant position in trading NYSE-listed shares, mostly to new electronic venues. It responded by changing its floor operation first into a hybrid market, and eventually into a fully electronic market. Also, HFTs were largely absent in 2001 but participated in about half of the trades at the end of the decade (SEC 2010). The quote-to-trade ratio increased tenfold, from 2 in 2001 to 20 in 2011 (Angel, Harris & Spatt 2015).

**Figure 2** shows that end-user transaction cost declined substantially in this period, by at least 50% in all categories. For example, the effective spread that investors paid on their market orders was 20 basis points (bps) in 2001 and 7 bps in 2011. This spread is a reasonable measure of the implicit trading cost for retail investors, as their orders are small enough to trade through a single market order. The explicit cost of trading for them (i.e., the broker commission) was 43 bps in 2001 and only 20 bps in 2011. Institutional investors trade larger orders and minimize their price impact by splitting them into into many small child orders that are fed to the market sequentially. Their total transaction cost, referred to as implementation shortfall, was 50 bps in 2001 and only 9 bps in 2011. The explicit cost for them was 15 bps in 2001 and 6 bps in 2011.

**Figure 2** further shows that US equity volume doubled from 2001 to 2011. This increase suggests that more securities were transferred from low- to high–marginal utility investors. I carefully picked the word "suggests," as trades are typically intermediated and therefore volume numbers overstate true reallocation among end-users. If intermediation chains lengthened with the entry of HFTs, then reallocations might not have grown as much as volume did (this is discussed further in Section 3.6).

The much lower transaction cost and strong volume growth do not, of course, prove that the current market structure is optimal. Perhaps electronic markets without HFTs could produce even better service at a lower cost. This is what the public debate has centered on (although these markets might not have arrived without HFTs, as argued above). Two popular books all but vilified HFTs, associating them with algo wars that leave "a path of destruction in their wake" (Patterson 2012, p. 63) or with "systemic market injustice" that arises from HFTs front-running orders (Lewis 2014, p. 70). Well-known financial economists have contributed by publicly sharing their impressions of high-frequency trading: Paul Krugman and Joseph Stiglitz, both Nobel laureates, expressed concern (Krugman 2009, Stiglitz 2014); Burton Malkiel emphasized benefit (Malkiel 2009); and John Cochrane shared how it all puzzles him (Cochrane 2012).

The public debate on HFTs has raged since 2009, thriving mostly on conjectures and fears. Opinions have failed to converge. The only thing all seem to agree on is that the key distinguishing feature of HFTs is their relative speed advantage for trading a particular security. One reason for such an advantage is information technology that enables them to quickly generate signal from the massive amount of (public) information that reaches the market every (milli)second. Examples are press releases by companies, governments, and central banks; analyst reports; and trade information, including order-book updates not only for the security of interest but also for correlated securities. Another, more narrow, reason for HFTs' speed advantage is quick access to such information (by using, for example, microwave towers; Laughlin, Aguirre & Grundfest 2014).

In parallel to this public debate, many scholars have thought deeply about how HFTs could affect market quality. Theorists have written down models to tease out what the social costs and benefits are of having extremely fast intermediaries in the market. Empiricists have creatively exploited the scarce data available.

The scope of this survey is the HFT literature. I review the theoretical papers on HFTs, weigh them by the amount of supportive evidence in empirical HFT papers, and come to a reasoned judgment on the net economic value of HFTs. This structure distinguishes this survey from the many excellent HFT surveys that have appeared thus far (Biais & Woolley 2011; Gomber et al.

2011; Chordia et al. 2013; Easley, López de Prado & O'Hara 2013; Jones 2013; Kirilenko & Lo 2013; Biais & Foucault 2014; Goldstein, Kumar & Graves 2014; SEC 2014; O'Hara 2015).

This survey is structured as follows. I start by discussing what can be learned about HFT by extrapolating from a classic model (Section 2). I then discuss the new models by grouping them into seven categories (Section 3). For each category, I also review the empirical papers that speak to either the assumptions of these models or their predictions. The survey concludes with an evidence-weighted verdict on whether HFTs benefit security markets or not. In this sense, we take stock of the HFT literature.

## 2. SPEED AND TRADING: EXTRAPOLATIONS FROM A CLASSIC MODEL

Before reviewing any new model proposed for HFTs, it is useful to ask what classic models can tell us about adding faster agents. Perhaps the model that speaks most to high-frequency trading is one that emphasizes how information asymmetry affects trading. A classic reference is the work of Glosten & Milgrom (1985), who model how a specialist issues price quotes to buy or sell a security: bid and ask quotes, respectively. He is risk-neutral and acts under competitive pressure. If information were symmetric across the specialist and the investors he trades with, then the zero profit condition would imply that he sets his bid and ask price equal to the expected payoff of the security. The bid–ask spread thus becomes zero, as he is not adversely selected on his quotes.

Asymmetric information, with some investors being better informed than others, drives a wedge between the bid and ask quotes. These informed investors may be thought of as corporate insiders or as "individuals who are particularly skillful in processing public information" (Glosten & Milgrom 1985, p. 77). I revisit this latter interpretation in the context of HFTs, in particular when discussing run games in Section 3.3.

This parsimonious model yields two important economic insights. First, the informed earn a positive profit effectively paid for by the uninformed. Second, the larger the information asymmetry is between the informed and the uninformed, the larger the transfers are between them. The cost for the uninformed might at some point exceed their private gains from trade, in which case information asymmetry leads to market breakdown.

How would speed affect trading in this model? If one were to replace the specialist by an HFT, then this speed upgrade would help the market maker maintain his price quotes. He is now able to quickly parse the public flow of information for value-relevant signals and to quickly update his quote when needed. This reduces information asymmetry, at least for informed investors who get their information advantage from public news. Loosely extrapolating from the Glosten–Milgrom model along these lines suggests the following market features:

- less adverse-selection cost;
- a consequently tighter bid-ask spread;
- more quote updates in between trades;
- more price discovery through these quote updates as opposed to trades;
- a higher trade probability as the tighter spread creates more trades either at the intensive margin (existing investors trade more) or at the extensive margin (new investors enter). The tighter spread enables investors to lock in private gains from trade that would otherwise have been too small to make up for paying the (half) spread.

These observations correspond to some of the empirical facts documented for HFTs, mostly reviewed in Section 3.1, which features models that embed the speed/informational friction of HFTs in the most common electronic market structure: limit-order books. The models echo some of these predictions developed by loose extrapolation, but derive them more rigorously in often much richer settings. For example, HFTs might endogenously become either markers or takers of price quotes.

I have discussed only one classic model to help shape intuition. For excellent reviews of classic theory, I refer the reader to O'Hara (1995); Biais, Glosten & Spatt (2005); Parlour & Seppi (2008); de Jong & Rindi (2009); Foucault, Pagano & Röell (2013); and Vayanos & Wang (2013).

**Other trader (OT):** any trader who is not an HFT

**High-frequency market maker (HFM):** an HFT who trades passively by submitting bid and ask quotes for others to take

**High-frequency bandit (HFB):** an HFT who trades aggressively by taking out stale quotes quickly after seeing (public) infomation

## 3. SPEED AND TRADING: INSIGHTS FROM NEW MODELS

The discussion of recent theoretical work on HFTs is grouped into seven subsections. Each of them reviews the economic insights that a new theory provides, discusses relevant empirical work on the issue, and concludes with a brief summary. All discussions are extremely brief. For example, I do not mention the data used in the empirical papers I review unless it is nonstandard (i.e., nonequity) data. Proceeding in this way guarantees a steady flow of insights and empirical facts.

The following terminology is used throughout. All traders are either HFTs or OTs (other traders). HFTs are further grouped into two types. The first type, HFMs (high-frequency market makers), trade passively by submitting bid and ask quotes. The second type, HFBs (high-frequency bandits), trade aggressively by taking out stale quotes quickly. **Table 1** characterizes all theoretical papers I review here in terms of their structure and results.

### 3.1. Speed Dispersion in Limit-Order Markets

Most electronic exchanges operate as limit-order markets. A limit order is essentially a price quote to either buy or sell a specified amount of securities. The exchange processes incoming orders in continuous time. It tries to match any new order against a stock of unexecuted orders. If there is such a match, then the incoming order is referred to as a market order. If in fact many matches are possible, then a market sell order is matched with the highest-priced limit buy, referred to as the best bid, and a transaction is concluded at that price. A market buy order trades with the lowest-priced limit sell, referred to as the best ask. If there is no match, the new order is added to the stock of unexecuted orders, the limit-order book. The distance between the best bid and the best ask is referred to as the bid-ask spread. I refer the reader to Parlour & Seppi (2008) for an in-depth review of limit-order markets.

**3.1.1. Theory.** The discussion of theory is organized into several subsections.

***3.1.1.1. HFTs as better-informed agents are good.*** Aït-Sahalia & Saglam (2014) analyze a more informed HFM in a continuous-time dynamic inventory model. They find that if the HFM is better able to predict the sign of future market orders, then the liquidity he provides improves. He adds more depth to his bid and ask quote (he cannot change the spread, as the authors assume it to be fixed). Goettler, Parlour & Rajan (2009) also find liquidity improvement when market makers become more informed about fundamental value. In their model, agents arrive randomly and, conditional on the state of the limit-order book, they decide to send either a limit or a market order. They find that low–private value agents endogenously become market makers. If these agents become more informed, liquidity improves.

***3.1.1.2. HFTs as faster-acting agents are bad.*** Bernales (2014) extends the model of Goettler, Parlour & Rajan (2009) by randomly assigning agents to be either fast or slow. He finds that

**Table 1 Structure/results of reviewed papers**

| Paper | Model structure | | | | | | Model results | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HFM endogenous? | HFB endogenous? | OT trading endogenous? | Information endogenous? | Market fragmentation? | Dynamic model? | Analytic results? | Effect on liquidity? | Effect on volatility? | Effect on welfare? | Policy evaluation?[a] |
| Aït-Sahalia & Saglam (2014) | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Baldauf & Mollner (2015) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ |
| Baruch & Glosten (2016) | ✓ | | | | | | ✓ | ✓ | | | |
| Bernales (2014) | ✓ | ✓ | ✓ | | | ✓ | | ✓ | ✓ | ✓ | ✓ |
| Biais, Foucault & Moinas (2015) | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | ✓ | ✓ |
| Bongaerts & Van Achter (2015) | ✓ | | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Bongaerts, Kong & Van Achter (2015) | ✓ | ✓ | | | | ✓ | ✓ | ✓ | | ✓ | |
| Boulatov, Bernhardt & Larionov (2016) | ✓ | | ✓ | | | ✓ | ✓ | ✓ | | | |
| Brunnermeier & Pedersen (2005) | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Budish, Cramton & Shim (2015) | ✓ | ✓ | | | | | ✓ | ✓ | | ✓ | ✓ |
| Cartea & Penalva (2012) | ✓ | | ✓ | | | | ✓ | ✓ | ✓ | ✓ | |
| Cespa & Vives (2015) | ✓ | | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Du & Zhu (2014) | ✓ | | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Foucault, Hombert & Roşu (2016) | | ✓ | | | | ✓ | ✓ | ✓ | | | |
| Foucault, Kozhan & Tham (2015) | ✓ | ✓ | ✓ | | ✓ | | ✓ | ✓ | | | |
| Glode & Opp (2016) | ✓ | | | | | | ✓ | ✓ | | ✓ | |
| Goettler, Parlour & Rajan (2009) | ✓ | ✓ | ✓ | ✓ | | ✓ | | ✓ | ✓ | ✓ | |
| Han, Khapko & Kyle (2014) | ✓ | | | | | | ✓ | ✓ | | | |
| Hoffmann (2014) | ✓ | ✓ | ✓ | | | ✓ | ✓ | | ✓ | ✓ | ✓ |
| Jarrow & Protter (2012) | | ✓ | | | | ✓ | ✓ | | ✓ | ✓ | |
| Jovanovic & Menkveld (2015b) | ✓ | | | | | | ✓ | ✓ | ✓ | ✓ | |
| Jovanovic & Menkveld (2015a) | ✓ | ✓ | ✓ | | | | ✓ | ✓ | | ✓ | |
| Li (2014) | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Li (2015) | | ✓ | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Menkveld & Yueshen (2014) | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | |
| Menkveld & Zoican (2016) | ✓ | ✓ | | | | | ✓ | ✓ | | ✓ | |
| Pagnotta & Philippon (2015) | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ |
| Roşu (2016) | | ✓ | | ✓ | | ✓ | ✓ | ✓ | ✓ | | |
| Weller (2013) | ✓ | | | | | ✓ | ✓ | ✓ | | ✓ | |
| Yang & Zhu (2015) | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | |

[a]Manuscripts in this column explicitly discuss regulatory proposals. Abbreviations: HFB, high-frequency bandit; HFM, high-frequency market maker; OT, other trader.

creating such speed dispersion leads to less gains from trade being realized. Slow traders are effectively forced out of using limit orders because of increased adverse selection risk. Hoffmann (2014) analyzes a model where HFTs quickly cancel their outstanding limit orders after news. This endogenously makes OTs post limit orders at less aggressive prices, reducing the trade rate and welfare. Du & Zhu (2014) vary speed in their model by adding more trading rounds per time unit. Only HFTs have the ability to keep up and be present in all rounds. They thus effectively force themselves in between trades of OTs, who do not have this ability. As a result, OTs experience higher trading cost.

***3.1.1.3. HFTs as faster-informed and faster-acting agents are either good or bad.*** Bongaerts & Van Achter (2015) focus on HFM price competition in a limit-order market and find that increased speed (in terms of contact rate) leads to faster undercutting and therefore benefits liquidity. If HFMs further have an informational advantage, then liquidity is reduced because of increased adverse selection risk for OT market makers. Jovanovic & Menkveld (2015b) find that the entry of well-informed and fast HFTs raises welfare when OTs use them to reduce information asymmetry between each other. An early OT seller might post an ask quote (through a limit sell order) but then suffer adverse-selection cost when trading with a late OT buyer who might have seen (common-value) news that arrived in between their arrivals. HFTs effectively offer the seller a cost-free route to the buyer when the seller sells to HFTs, who can post and update an ask quote aimed at the late buyer. HFTs have the information and speed to do such updating. Jovanovic & Menkveld show that more gains from trade are realized this way. If, however, HFTs see news that even the buyer in this model is not aware of, then they can reduce gains from trade, as they effectively add adverse-selection cost to the trading game. An important implication that follows from this analysis is that one needs to exercise extreme caution when interpreting lower bid-ask spreads as better market quality. In the model, a reduced spread owing to HFT entry might be dominated by OTs not being able to earn the spread by posting a limit order themselves (because of the threat of adverse selection by HFTs).

**3.1.2. Evidence.** The discussion of evidence is organized into several subsections.

***3.1.2.1. HFTs are extremely fast.*** Menkveld (2016) analyzes nanosecond data and finds that 20% of trades cluster within a millisecond. The analysis suggests that HFT response times are on the order of microseconds (one microsecond is $10^{-6}$ s).

***3.1.2.2. High-frequency market making.*** The early empirical papers examined algorithmic traders (ATs), a group that includes HFTs but also, more generally, all traders who use computers to automatically make trade decisions. An example of a non-HFT AT is one who aims to minimize the price impact of a large order that has to be executed. Such an order is typically executed through a series of smaller child orders that are sent to the market sequentially.

Hendershott, Jones & Menkveld (2011) exploit an exogenous NYSE event to establish that ATs causally reduce the bid-ask spread. They find that price quotes experience lower adverse-selection cost, which explains the lower spread. They further document that more algorithmic trading results in more price discovery through quotes as opposed to trades. Malinova, Park & Riordan (2013) also find, by exploiting a Canadian regulatory fee change, that ATs reduce spread.

Hendershott & Riordan (2013) show that ATs are quick to post limit orders when spreads are wide, but take limit orders when spreads are narrow. Latza, Marsh & Payne (2014) focus on fast market orders. They define fast market orders as those that are quick to arrive after a limit order is added to the book (they arrive within 50 ms after, and trade with, these limit orders). The authors find these fast market orders to be uninformed, potentially indicating HFMs offloading inventory.

Empirical studies on data samples with HFT identification generally echo the findings of AT studies. In particular, HFTs are more likely to add limit orders to the book when the spread is wide, and thus supply liquidity (Carrion 2013; Hagströmer, Nordén & Zhang 2014; Jarnecic & Snape 2014). Similarly, Yao & Ye (2015) find that HFT liquidity supply is larger for stocks for which the spread is constrained to be large because of tick size.

More HFT competition seems to further decrease the spread and to lead to quicker recovery of the spread after it is widened (e.g., after a market order arrival). Hasbrouck (2015) finds positive skewness for ask quote changes and negative skewness for bid quote changes. He interprets this finding as Edgeworth cycles: HFTs undercut each other, leading to gradual price improvement until a large market order hits and removes the best price quote(s), thus making the best quote move, potentially by multiple ticks. Brogaard & Garriott (2015) find that HFT entry is followed by a reduction in the bid-ask spread. The result is strongest when an HFT is the first to enter.

HFTs further supply liquidity by trading against (transitory) price pressures. Boehmer, Li & Saar (2015) find that HFTs follow similar strategies, and the extent of their competition (or correlated activity) is negatively related to short-term volatility. They interpret this finding as consistent with competition between HFTs in market-making. Brogaard, Hendershott & Riordan (2014) use the state space model of Hendershott & Menkveld (2014) to identify the permanent and transitory components in price changes. They find that HFTs trade against transitory shocks. Brogaard et al. (2015a) analyze extreme price movements and find them to be caused by OT order imbalance. HFTs trade against this imbalance and thus stabilize prices. Benos et al. (2015) also find a strong commonality in HFT order flow, and HFT trading "does not generally contribute to undue price pressure and price dislocations" (Benos et al. 2015, p. 1).

Colocation events provide what is perhaps the best laboratory to study how HFTs affect limit-order markets. A reasonable conjecture is that among all traders, HFTs benefit the most from colocation, the service an exchange provides (at a fee) to locate close to the exchange server. One could therefore interpret the results of a colocation event study as the outcome of more HFT activity. Boehmer, Fong & Wu (2014) find that colocation reduces the bid–ask spread and improves the quality of price discovery. Gai, Yao & Ye (2013) also find that order book liquidity improves, but through more depth rather than an improved spread. Brogaard et al. (2015b) study a colocation event in which traders are given various options. They find that the fastest colocation service is primarily used for HFT market making. This suggests that HFTs endogenously become market makers, as predicted by theory.

***3.1.2.3. Type of information.*** The evidence discussed thus far shows that HFTs are quick to snap up the bargains provided by spread-improving limit orders, but are they also more informed? If so, what type of information do they have access to? To whet the reader's appetite, **Figure 3** illustrates that an HFM's limit orders (right-hand side) are relatively short-lived and generally in sync with the best bid and ask quotes when compared to all other limit orders (left-hand side). This pattern suggests quoting by a fast and informed trader.

HFTs seem to actively manage their limit orders, and these quote updates reflect information. Hasbrouck & Saar (2013) provide the most detailed account of active limit-order management. They find that about 60% of quote cancellations are followed by resubmissions within 100 ms (of which 49% are at 1 or 0 ms). The median stock experienced 26,862 limit-order submissions (daily average), 24,015 limit-order cancellations, and 2,482 market orders. Brogaard, Hendershott & Riordan (2015) document that the majority of price discovery occurs through quote updates as opposed to trades. HFTs produce most (60–80%) of this information in quote updates.

a b

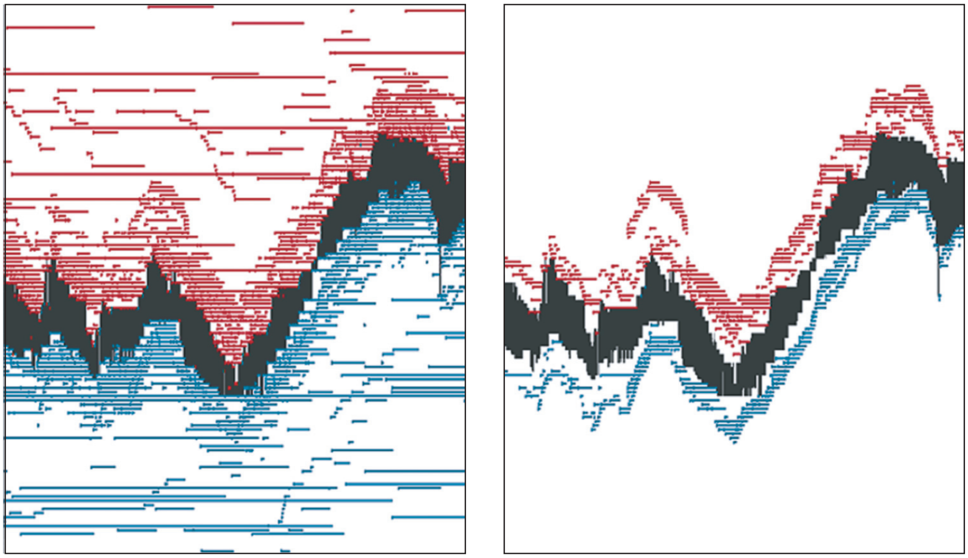

**Figure 3**

Book dynamics for one security traded at Euronext Amsterdam. Panel *a* illustrates dynamics of the entire book, where time is on the horizontal axis and price is on the vertical axis. A red/blue bar depicts the life of a limit sell/buy order (ask/bid quote). The black area depicts the bid-ask spread. Panel *b* repeats this graph but only shows the limit orders of one high-frequency trader. Figure adapted with permission from AFM (2016, figure 1).

HFTs are quick to process public information. Hu, Pan & Wang (2014) show how powerful their textual algorithms are when parsing macro announcements. They find that when news is released to HFTs 2 s before the official announcement, most of the price discovery of index futures happens within 0.2 s after HFTs had their early peek. HFTs also parse order flow to predict price changes. Harris & Saad (2014) show that future returns can indeed be predicted based on the (publicly available) message traffic between traders and the exchange. The size of message traffic quickly makes this a big data challenge. HFTs further try to obtain such data quickly, typically by subscribing to proprietary data feeds made available by the exchange. Ding, Hanna & Hendershott (2014) show that these feeds lead the public data feed by milliseconds. Finally, Jovanovic & Menkveld (2015a) find that a large HFM uses index futures information to update his quotes.

HFMs are informed, but cannot completely avoid being adversely selected on their quotes, as found by Menkveld (2013) and Brogaard, Hendershott & Riordan (2014). Fishe, Haynes & Onur (2015) study trading around local price trends. They find that those who best predict either the start or the end of such trends are typically not the fastest traders.

*3.1.2.4. Net effect of HFT entry in limit-order markets.* The evidence suggests that HFTs are both fast and informed. Theoretical studies predict results to be mixed in this case. Jovanovic & Menkveld (2015a) calibrate their model based on an event that led to HFT entry. They find a moderately positive welfare effect.

**3.1.3. Summary.** HFTs reduce market quality if they are faster to act than others. They increase market quality if they are better informed. If both, then the prediction is nontrivial. The evidence is that they are both extremely fast to act and better informed. Colocation event studies and a model calibration suggest that the net effect for market quality is moderately positive.

## 3.2. Speed Used to Prey on Large Orders

**3.2.1. Theory.** The discussion of theory is organized into several subsections.

***3.2.1.1. HFTs preying on large uninformed orders.*** Brunnermeier & Pedersen (2005) show that strategic traders could prey on (inelastic) demand from a distressed uninformed trader. This trader sees liquidity suddenly diminish when predatory traders trade in the same direction. This paper is the only non-HFT paper added to this review, simply because HFTs are often referred to as predatory traders. They dominate the market in the short run and rationally exploit their speed.

***3.2.1.2. HFTs preying on large informed orders.*** HFTs will prey on informed orders once they learn about them, thus delaying price discovery and potentially reducing long-term efficiency. Li (2014) finds that if HFTs anticipate incoming orders, then liquidity is reduced, especially when these HFTs have heterogeneous speeds. Yang & Zhu (2015) find that if HFTs become particularly good at detecting informed orders, then their back-running will make the informed trader switch to a mixed strategy. Both Li and Yang & Zhu emphasize that the informed trader's endogenous response of trading less aggressively delays price discovery. The trader might forgo collecting information altogether, in which case this information might never be revealed in prices, thus hurting long-term efficiency.

Boulatov, Bernhardt & Larionov (2016) study Nash equilibria when multiple traders seek to minimize transaction cost when trading continuously in a fixed time period. Each trader takes the price impact function as exogenous. This function makes the net order flow have both a transitory and a permanent price impact (in the sense of Almgren & Chriss 2000). They find that traders who have a small position to trade either lean against the orders of large-position traders or prey on them. Both types of behavior can occur in equilibrium. The large-position traders' optimal response to preying is to delay trading.

**3.2.2. Evidence.** The findings in various empirical studies suggest that HFTs are able to predict order flow. Hirschey (2016) provides perhaps the most direct evidence. He reports that HFT aggressive flow at a given second predicts OT aggressive flow in the following 30 s. He interprets this finding as anticipatory trading by HFTs. In further analysis, he discards various alternative interpretations: HFTs reacting faster to news, positive-feedback trading by OTs, or HFTs being misclassified as OTs. Breckenfelder (2013) finds that HFT entry makes stocks less liquid as measured at a daily frequency. A new HFT might have methods of detecting informed orders that differ from the methods of incumbent HFTs, thus worsening trade conditions for OTs. Raman, Robe & Yadav (2014) find that electronic market makers withdraw at times of directional flow.

Some studies focus on institutional investors executing large orders through multiple (child) trades. They find that transaction costs are higher when HFTs run on their orders. Brogaard et al. (2014) use exchange speed upgrades to identify a causal effect. They find that HFT activity increases around such events, but institutional investor cost remains unchanged. They are careful to note that statistical power is low, as HFTs are only slightly more active and transaction costs of institutional investors are volatile. Korajczyk & Murphy (2016) and van Kervel & Menkveld (2015) analyze HFT trading in the lifetime of institutional orders. They both document that HFTs initially lean against large institutional orders (e.g., they buy when the investor sells) but eventually switch and trade along with the long-lasting ones. Korajczyk & Murphy (2016) focus on HFT market making and inventory management. Van Kervel & Menkveld (2015) further analyze the same-direction trading. They find that it is stronger than simply closing the initial position built up when trading against the institutional order. HFTs not only close the position but seem to enter a position in the direction of the order. Van Kervel & Menkveld show that such behavior is profitable for HFTs and that, coincidentally, transaction costs are higher for institutional investors when HFTs engage in such trading. They further show that the institutional orders for which this happens are informed and interpret the evidence as being consistent with back-running.

Finally, less price discovery ahead of announcements might be early evidence of delayed price discovery stemming from predatory trading. Weller (2016) finds that algorithmic trading proxies correlate positively with price jumps on earnings announcements. This finding suggests that AT presence discourages costly information acquisition ahead of an announcement.

**3.2.3. Summary.** If HFTs have the ability to detect large orders of OTs, then they can prey on them. This raises OT transaction cost and leads to slower price discovery. The evidence suggests HFTs are able to predict OT orders and to back-run on long-lasting informed orders of institutional investors. The institutional investors experience higher transaction costs when this happens.

## 3.3. Speed in a Run Game After (Public) News

**3.3.1. Theory.** HFTs all trading on the same public signal creates costly run games in continuous-time markets. That public-news trading comes with a negative externality was well known for human-intermediated markets (Foucault, Röell & Sandås 2003). Budish, Cramton & Shim (2015) explore this phenomenon for high-frequency trading in continuous-time markets and show that a wasteful arms race becomes inevitable. When news breaks, both HFMs and HFBs run to the market. If an HFM arrives first, he updates his outstanding quote. If an HFB arrives first, he trades with the stale quote and thus imposes adverse selection on the HFM. The HFM anticipates such a cost and raises the spread accordingly, effectively making OTs pay the cost. This run game incentivizes HFTs to engage in a wasteful arms race (e.g., by building towers, as shown in **Figure 4**). Budish, Cramton & Shim argue that the root cause of this dead-weight cost is that venues match orders in continuous time. They show that changing to (high-frequency) discrete time batch auction reduces the benefit of speed investment by an order of magnitude. This largely avoids the costly arms race.

***3.3.1.1. Run games in the context of multiple markets.*** Costly speed races seem particularly relevant in multimarket settings, where an event in one market (e.g., the arrival of a market sell order) serves as the public signal. Foucault, Kozhan & Tham (2015) show that such a run game creates toxic arbitrage and thus widens the bid–ask spread. They essentially add speed as a choice variable to earlier work on arbitrage and differential speed (Kumar & Seppi 1994). Biais, Foucault & Moinas (2015) argue that becoming an HFB (a fast trader) is privately beneficial, as it enables one to find the best price in a fragmented setting, yet it is socially costly, as it creates adverse-selection cost for low-frequency market makers. Baldauf & Mollner (2015) show that increasing exchange speed stimulates cross-market HFB activity and thus reduces the incentive of an analyst to collect information. The moment an analyst order hits a market, HFBs learn about it and quickly trade on this signal in other markets. Random communication latency (delay) makes it impossible for all analyst orders to arrive at all markets at exactly the same time. This reduced ability for an analyst

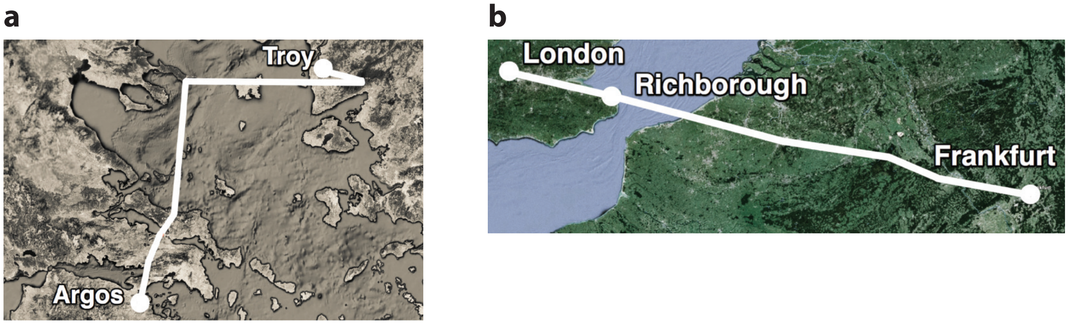


**Figure 4**

Optical data transmission, then and now. Light has been used for fast communication ever since the classical age. (*a*) Troy–Argos, 1185 BC: Agamemnon used prearranged beacon fires to announce the fall of Troy to his wife Clytemnestra in Argos. Speed turned against him, as it gave her and her lover time to prepare for his assassination. Note that the small detour at Troy adds latency to the 1-bit communication line: Light needed to travel up the nearest mountain before it could travel across the sea. Panel *a* adapted with permission from Laughlin (2014). (*b*) Frankfurt–London, 2016 AD: Vigilant Global, a high-frequency trader, reportedly plans to build a tower taller than the Eiffel Tower in Richborough, a town in southeast England. In a local town hall meeting, a spokesman addressed worried residents but was vague about what the tower would be used for. Panel *b* adapted with permission from Laumonier (2016).

to trade on his information harms long-term efficiency. [This idea is similar in spirit to that of Li (2014) and Yang & Zhu (2015), discussed in Section 3.2.1.]

***3.3.1.2. Run games with speed heterogeneity across HFTs.*** HFTs' operating at different speeds has a nontrivial effect on the bid–ask spread. Han, Khapko & Kyle (2014) show that informed fast market makers impose a winner's curse on slow ones. A crucial parameter is the probability that fast HFMs show up in the market (an event that is independent of news). If fast HFMs appear with higher probability, then the ever-present slow HFMs have to raise the spread they charge. However, these wider spreads are more likely to be undercut by fast HFMs who arrive with higher probability. The authors show that the net effect could go either way. Foucault, Hombert & Roşu (2016) propose a dynamic model where securities with more informative news attract more HFBs, yet the average effective spread is lower. The reason is that trading with HFBs after news imposes an adverse-selection cost on market makers, thus raising the spread, but the more informative news itself lowers the spread. The latter effect dominates in equilibrium.

***3.3.1.3. The effect of exchange speed on run games.*** A faster exchange could either lower the bid–ask spread or increase it. Menkveld & Zoican (2016) find that making exchanges faster raises the probability of an HFB–HFM run game after news. A slower exchange helps an HFM to avoid such duels because, at the moment news hits, the exchange might still be processing an OT order that is on its way to the matching engine. If this is the case, the outstanding HFM quote will meet the OT order and thus evade the costly HFT duel. The upside of a faster exchange is that it allows an HFM to refresh his quotes more often and thus reduce adverse-selection cost, should it come to a duel. Menkveld & Zoican find that increasing exchange speed raises the bid–ask spread if the ratio of news to liquidity traders is low but lowers it when this ratio is high. Li (2015) analyzes exchange speed by changing the frequency of batch auctions. He generalizes the standard model of Kyle (1985) by replacing the monopolistic informed trader by an oligopoly of them. He finds that batching orders less frequently does not necessarily improve market quality. The profitability of informed traders drops, and some might have to exit for informed trading to remain profitable (net of a fixed cost). The ones who remain experience less competition and extract more rents, and this increases the transaction cost of others.

***3.3.1.4. Persistent pricing errors.*** Jarrow & Protter (2012) show that if HFTs collectively trade on public signals, this does not necessarily yield "value discovery." They show that it might lead to persistent pricing errors in such a way that the no-arbitrage condition is not violated.

**3.3.2. Evidence.** HFTs actively engage both in market-making and in sniping stale quotes. Hagströmer & Nordén (2013) document strong persistence in types, with about half a dozen HFM types and two dozen HFB types. Benos & Sagade (2016) find the same and further report that both types seem to be informed. Brogaard, Hendershott & Riordan (2014) carefully distinguish between transitory and permanent price changes and find that HFT market orders trade in the direction of permanent price changes, and HFM limit orders opposite to them. Brogaard, Hendershott & Riordan (2016) use the 2008 short sales as an instrument to identify that HFTs adversely select limit orders and thus affect liquidity negatively. The evidence in these papers is consistent with the hypothesized HFB–HFM duels.

Other studies provide perhaps more direct evidence of HFB–HFM run games. They show that a stronger HFB presence correlates with a higher spread, with only the fastest market makers still earning a positive (gross) profit. Brogaard & Garriott (2015), for example, find that the exit of aggressive HFTs leads to an improved spread. Van Kervel (2015) documents that an increase in fast traders imposes additional cost on market makers who quote in multiple venues. Biais, DeClerck & Moinas (2015) find that only the fastest traders earn a small positive (gross) profit on their limit-order executions; the adverse-selection cost they incurred was just short of the (half) spread they earned. Menkveld (2016) finds the same for a dataset with HFT identification: HFTs earned money on their limit-order executions and OTs lost money on limit-order executions (net of the rebate that market makers received in case their order executed; there was no such rebate in the sample of Biais, DeClerck & Moinas). Chaboud, Hjalmarsson & Vega (2015) study the introduction of a 250-ms minimum quote life and find that it reduced the adverse-selection cost for OTs, but did not change the bid–ask spread. These two observations are not necessarily inconsistent; Han, Khapko & Kyle (2014) emphasize that the lower spreads that OTs are able to post might be offset by fewer HFMs willing to undercut them (as they cannot quickly remove their quotes after seeing news). The two effects might cancel so that the average spread remains unchanged.

The multimarket models receive some empirical support from two foreign exchange studies. They both find that HFBs are highly active in a multimarket setting. Chaboud et al. (2014) analyze triangular arbitrage and document that arbitrage opportunities decline after machines (HFBs) enter the market. Foucault, Kozhan & Tham (2015) find that the days with more run games (toxic arbitrage) are also the days with higher spreads.

Finally, the evidence on how an exchange speed upgrade affects the bid–ask spread suggests a weakly negative correlation. Riordan & Storkenmaier (2012) find a lower spread after a speed upgrade, but only for small stocks. Gai, Yao & Ye (2013) find no relationship between the two.

**3.3.3. Summary.** Public signals trigger HFT run games. HFMs run to update their outstanding quotes; HFBs run to take out stale quotes. These run games make trading more costly for OTs, in particular because HFTs need to recoup the costs of an arms race. Exchange speed affects these run games in nontrivial ways. The evidence is largely consistent with run games. It is mixed regarding how exchange speed affects the bid–ask spread, as predicted by theory.

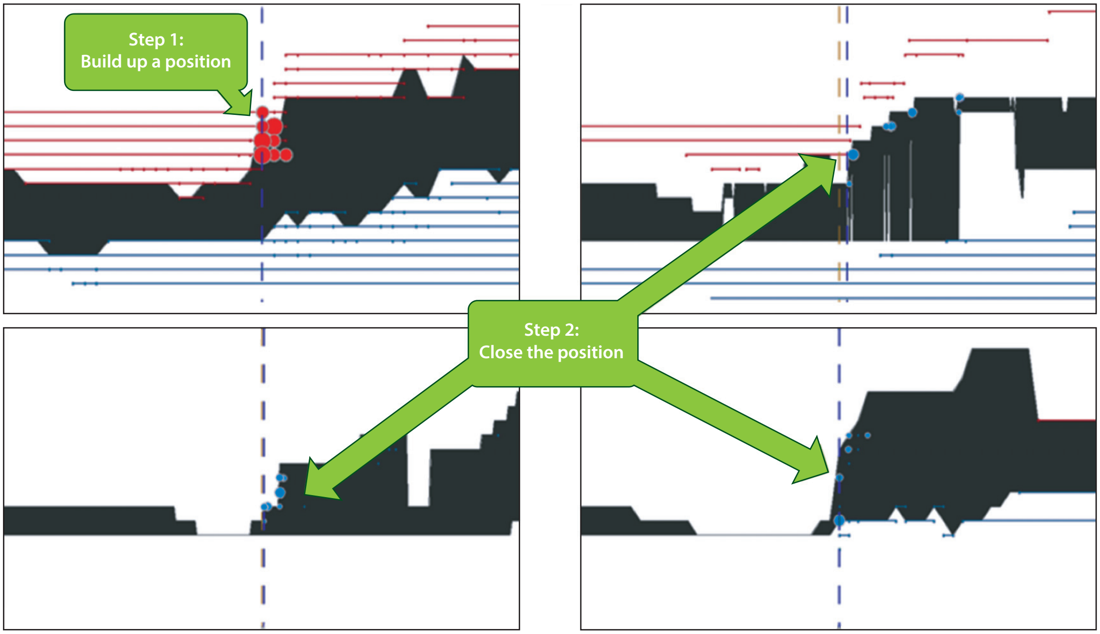


**Figure 5**

How one HFT (high-frequency trader) traded in a large Dutch cap in four limit-order books. The red/blue bars depict the life of his limit sell/buy orders (ask/bid quotes). The red/blue dots represent his sell/buy transactions. Larger dots correspond to larger transactions. The black area depicts the bid-ask spread. The top right graph shows how some traders take the HFT's ask quotes. The other graphs show how the HFT quickly closes his position by taking the ask quotes of others in the three other limit-order books. Figure adapted with permission from AFM (2016, figure 8).

## 3.4. Speed to Connect Fragmented Markets

Menkveld (2014) argues that the two most salient trends in securities markets since the turn of the century—order flow fragmentation and HFT entry—are intimately related, and that both are driven by technology and regulation. Consider that it is difficult for new exchanges to enter in human-intermediated markets. Opening up another trading floor and charging a lower fee is unlikely to attract order flow. The reason is that search costs (checking for a better price quote at the other floor) are high, and therefore the liquidity externality is extremely high in these markets; traders go where other traders are. The incumbent exchange benefits, as the competitive threat is low.

Technology and regulation changed this status quo, most dramatically in the first decade of the century. Trading floors were replaced by exchange servers matching electronic buy and sell messages. New regulation, Reg-NMS in the United States and MiFID in Europe, allowed for the entry of new trading venues. Entry also became more viable as search costs declined by an order of magnitude when humans on telephones were replaced by computers on networks. It became trivial to poll other new venues for better price quotes.

HFTs specialized in cross-market arbitrage and effectively matched buyers and sellers across venues. **Figure 5** illustrates this by showing how one HFT traded a large-cap stock across four limit-order books. The moment his ask quotes were being taken in market A, he ran to markets B, C, and D to take the ask quotes of others and thus rid himself of the position. He connected buyers in A with sellers in B, C, and D. It makes economic sense that one group should specialize in this type of activity, as only they then pay the nontrivial cost of maintaining high-speed connectivity to all markets.

Venues' competition for order flow benefited end-users in several ways. First, it created a downward pressure on trading fees. Second, it strengthened competition among limit-order submitters, leading to higher overall depth (Foucault & Menkveld 2008). Third, it fostered innovation and created heterogeneity in venues. Such heterogeneity either could improve liquidity in the presence of homogeneous investors (Parlour & Seppi 2003), or could create value simply by serving investors with different needs.

Two recent theoretical studies both identify economic costs of speed and fragmentation. They therefore provide some counterweight to the rather rosy perspective sketched thus far.

**3.4.1. Theory.** Pagnotta & Philippon (2015) find that exchange competition lowers fees and raises investor participation, but that entry could be excessive. Exchanges position themselves strategically with respect to speed, with some exchanges paying for a faster system to serve investors with highly volatile private values. The fastest type of exchange over-invests in speed from a social perspective, as it privately benefits by differentiating itself from other exchanges, and this lowers fee competition (by effectively creating a captive client base). The slowest type under-invests for the same reason.

Menkveld & Yueshen (2014) find that fragmented markets benefit from HFT presence, except when reselling opportunities suddenly turn out to be low (e.g., because of reduced intermarket connectivity). As an illustration, HFTs might have just taken on inventory by buying from a large seller when they privately learn that reselling opportunities are low. They quickly trade the hot potato among themselves at a rapidly declining (endogenous) price. The large seller interprets this volume increase as possibly indicating that many other end-users are present and rationally decides to accelerate selling. The authors argue that their stylized rational-expectations model could explain the Flash Crash.

**3.4.2. Evidence.** Menkveld (2013) finds that exchange entry reduced the bid–ask spread, but only after a large HFT began trading on both the incumbent and the entrant exchanges. The HFT is shown to have been equally active in both exchanges. Malinova & Park (2016) document how these modern market makers operate across markets.

The Flash Crash evidence is largely consistent with suddenly reduced reselling opportunities that left HFTs trading a hot potato in E-Mini. Kirilenko et al. (2014, figure 8) show that the steepest price drop coincides with a steep increase in HFT participation in trades. Menkveld & Yueshen (2015) find that cross-market arbitrage broke down one minute before the crash. This, in turn, might have been the result of data feeds suddenly slowing down, as discussed in Aldrich, Grundfest & Laughlin (2016). Some OTs responded by submitting inter-market sweep orders. The stocks that fell most during the Flash Crash experienced the most inter-market sweep orders (McInish, Upson & Wood 2013). This could explain why the most fragmented exchange-traded funds suffered the deepest price declines (Madhavan 2012). Strongly fragmented securities rely most on cross-market arbitrage and become vulnerable once such arbitrage disappears. I refer the reader to SEC (2010) for a detailed description of the Flash Crash.

**3.4.3. Summary.** HFTs and new venues have needed each other to thrive; HFTs have needed lower fees and new venues have needed attractive prices quotes. Investors benefit from lower fees and more innovation. Recent theory suggests that venue competition could be further improved and that (rare) Flash Crash events exist in equilibrium. The evidence is largely in line with these predictions.

## 3.5. Speed as a Source of (Endogenous) Quote Flickering

**3.5.1. Theory.** Quote flickering—the rapid change of price quotes—could be the result of healthy HFT competition. Prices of others are easily learned in electronic limit-order markets, and one's own price quote is easily changed. Baruch & Glosten (2016) show that fleeting orders could arise in an electronic limit-order market because of HFMs playing a mixed strategy over prices. They further show that HFTs could earn rents when there are only finitely many HFTs, but that these rents disappear when the number of HFTs goes to infinity.

Jovanovic & Menkveld (2015a) find that a pure-strategy equilibrium does not exist when HFMs pay a (tiny) cost to submit a price quote. They show that in this case there is a unique mixed-strategy equilibrium. HFMs first toss a (biased) coin to decide on whether or not to issue a price quote. If yes, they then randomly pick a price quote from a nondegenerate distribution. Contrary to the findings of Baruch & Glosten (2016), Jovanovic & Menkveld find that any HFT added beyond the first two raises the cost for OTs. The reason is that more HFTs competing adds to the total dead-weight cost of submitting price quotes. This is ultimately the result of HFTs not internalizing the negative cost they impose on others when deciding to submit a price quote.

**3.5.2. Evidence.** High-frequency flickering of quotes is episodic in nature. Hasbrouck (2015) argues that regular quote changes are likely the result of HFTs undercutting each other after market orders remove price quotes from the book, as discussed in Section 3.1.2.2. High-frequency flickering is likely to explain why Egginton, Van Ness & Van Ness (2016) find episodic bursts of quoting activity not triggered by market orders. They report that 74% of US securities experience such spikes at least once per year. Cartea et al. (2016) study a 2007–2015 sample and find that the number of quotes that were canceled within 100 ms peaked at, on average, 61.8 per minute in 2014.

Price quotes seem to have become noisier over time. Hasbrouck (2015, table X) analyzes variance ratios. He shows that 256 times the variance of 50-ms quote returns was a factor of 2.90 larger than the variance of $(256 \times 50)$-ms (12.8-s) quote returns in 2001. This ratio suggests that there is substantial noise in prices; it far exceeds 1, which is the value for a random walk benchmark. He further shows that this ratio increased over time to 3.16 in 2011, an increase of 9%. If extreme peaks are clipped, then this ratio almost doubled from 1.55 in 2001 to 2.91 in 2011. This shows that the nature of quote changes changed over time; in 2011 they exhibited relatively more oscillations of lower amplitude. This might be because of more flickering and fewer Edgeworth cycles over time. Yueshen (2015) uses a reduced-form structural model to estimate the time trend in excess price dispersion. He estimates the size of noise in relative terms, expressed as a fraction of information flow. He finds that for second-by-second quote returns in 2001, the noise is about twice the size of the information in order flow (Yueshen 2015, figure 8). In 2011 this ratio more than doubled, with noise being about five times larger.

**3.5.3. Summary.** Quote flickering seems endemic to HFMs competing on price. They resort to random quoting (mixed strategies) to avoid being undercut. The evidence suggests that quote flickering is episodic in nature. It seems to have grown over time, in particular in the 2001–2011 period of quick migration to electronic venues and rapid HFT growth.

## 3.6. Speed to Create Productive Intermediation Chains

**3.6.1. Theory.** HFTs increase the transaction cost for others if they only add a layer of intermediation. Cartea & Penalva (2012) formalize this conjecture by analyzing HFTs that interpose themselves between liquidity traders and professional traders. These HFTs exploit their monopoly to extract part of the trade surplus and thus raise the cost of trading for OTs.

In three papers, researchers argue that a longer intermediation chain could actually benefit liquidity. First, Weller (2013) shows that a chain that results from the fact that not all HFMs are equally fast increases liquidity supply. This is because the speed heterogeneity creates a productive ordering of events. The fastest HFMs pick out the most attractive (i.e., least informed) market orders, quickly cancelling ahead of unattractive orders. These unattractive orders thus land on slower HFMs, causing them to experience higher (adverse-selection) cost. This additional cost might, however, be more than offset by these last movers having a lower inventory to begin with (for the simple reason that they are last movers). This implies that their marginal cost for taking one more security into their inventory is lower. The author cautions that endogenizing speed might prevent such potentially productive sequencing.

Second, Glode & Opp (2016) argue that intermediation chains could benefit liquidity, as such chains enable intermediaries to sequentially parcel out the burden of information asymmetry across themselves. It would be too costly for any individual intermediary to intermediate alone, as he would have to charge a price higher than the gains from trade, thus shutting down the market.

Third, Roşu (2016) relates hot potato trading to an intermediation chain effect in a model with symmetrically and privately informed HFBs. He argues that if some HFBs are faster than others, then they trade more aggressively than in a benchmark case where all HFBs have the same speed. This enables a market maker to learn more quickly and charge a lower spread on average (measured as the price impact of a market order). The author cautions that the model only focuses on HFBs, not HFMs, as the market makers are assumed to be slow.

**3.6.2. Evidence.** Weller (2013) presents direct evidence on intermediation chains in commodity futures trading. High-frequency market makers provide rapid execution to investors and effectively consume inventory risk-bearing services from slower market makers. He finds that 65–75% of all trades are intermediated. The median chain is two members long, and the longest chains (at the 90th percentile) consist of six members.

Speed not only sets HFTs apart from OTs but also sorts HFTs into fast and fastest. Baron et al. (2015) find that speed is a meaningful attribute of an HFT. Some HFMs are persistently faster than others, and the same goes for HFBs. Baron et al. further find that the fastest HFTs earn higher profits. An exchange fosters such speed heterogeneity by offering a menu of colocation services, with faster ones being more expensive. Brogaard et al. (2015b) study trading before and after an exchange offered a richer menu of colocation services. If exchanges strive for maximum trade among investors, then such encouragement of speed heterogeneity is evidence that they believe intermediation chains are good for trade. The authors find that HFTs sorted themselves across the various options (not all bought the fastest service). They further show that the bid–ask spread declined and that depth improved after the event, consistent with intermediation chains benefiting liquidity.

**3.6.3. Summary.** Intermediation chains can be useful in completing trade between end-users, by forcing intermediaries to line up in a productive sequence; by having intermediaries effectively share the burden of an information asymmetry; or by having faster intermediaries trade more aggressively, thus revealing information early and reducing information asymmetry later. The evidence shows that intermediation chains exist, with some intermediaries being persistently faster than others. More speed heterogeneity among HFTs is shown to reduce spread and raise depth.

## 3.7. Speed to Satisfy Investors' Appetite for Faster Trade Completion

**3.7.1. Theory.** If investors appreciate faster trade completion, then an HFT arms race could be desirable from a social perspective. Bongaerts, Kong & Van Achter (2015) formalize this argument in a standard Merton (1969) model. They first observe that, in a continuous-time setting, price changes generate continuous-time rebalancing needs for an investor with constant relative risk aversion. They prove that for such an investor, the value of rebalancing increases less than proportionally with speed in the limit. If technology cost increases more than proportionally with speed, then an arms race is wasteful in the limit. More generally, they note that the net result critically depends on investor preference (i.e., the shape of the utility function).

Investors may appreciate extremely fast trading but should expect episodic flash crashes to come with it. Cespa & Vives (2015) find that if some investors' hedging needs are extremely short-lived, to the point that only some intermediaries are in synchronization, then market freezes could occur in periods when not all intermediaries are present. This is because the absence of some intermediaries enlarges order imbalance, potentially to the point of market breakdown.

**3.7.2. Evidence.** The evidence is thin on this issue. Empiricists have not converged on the shape of investor utility functions, thus complicating judgment on how utility scales with trading speed. Moallemi & Sağlam (2013) approach the problem from a completely different angle to get an estimate of how beneficial lower latency is for investors. They calibrate the solution of an optimal-control problem in a partial equilibrium setting. Lower latency essentially allows an investor to reduce adverse-selection cost on his price quotes. Their calibration suggests that reducing latency from a human reaction time of 500 ms to a machine reaction time of 5 ms reduces such adverse-selection cost from 20% of the bid–ask spread to 5%. Finally, the prediction by Cespa & Vives (2015) of occasional flash crashes seems to be borne out by Easley, López de Prado & O'Hara (2011, 2012), who find that (initiator-signed) order imbalance was extremely high just before the Flash Crash.

**3.7.3. Summary.** Faster trade completion might benefit investors, as it keeps them closer to the frictionless target portfolio. If, however, markets tick too fast for intermediaries to keep up, it might come with occasional market freezes (flash crashes). The evidence on these issues is thin.

## SUMMARY POINTS

1. Transaction costs have decreased substantially. In the decade of migration to electronic trading and HFT arrival, transaction cost decreased by over 50% for both retail and institutional investors.
2. HFT market-making reduces transaction cost. If HFTs enter limit-order markets with only a speed advantage, they hurt trading. If they enter with an informational advantage, they benefit trading by endogenously becoming market makers. If both, then the outcome depends on the strength of both forces. The evidence is that HFTs are extremely fast and well informed. As predicted, they are important price quote submitters (i.e., market makers). A calibration finds HFT entry to have a modest positive welfare effect.
3. HFT preying on large orders increases transaction cost. If HFTs have the ability to predict institutional order flow, then they might extract rents by preying on these orders. The evidence suggests that HFTs are mostly market makers for large institutional orders, and only prey on them when they are extremely large and execute through a long series of child trades.

4. An HFT run game on public signals increases transaction cost. In continuous-time markets, HFTs (both those with quotes outstanding and those without) race to the market on a public signal. Market makers run to update their stale quotes; bandits run to take them out. This run game comes with an arms race that raises transaction cost. Exchange redesign could reduce this cost.
5. HFTs facilitate venue competition. Cross-market arbitrage by HFTs effectively connects buyers and sellers across venues. HFTs thus make venue competition possible, with more innovation and lower trading fees as a result.
6. Quote flickering may be a by-product of healthy HFT competition. Quotes flicker when HFTs play mixed strategies in equilibrium. HFTs resubmit randomized price quotes in order to avoid being undercut. The evidence is indicative at best. Quote uncertainty has increased over time and is episodic in nature.
7. HFTs may operate in productive intermediation chains. An HFT intermediation chain could reduce information asymmetry and thus benefit trade. The evidence suggests such chains are prevalent.
8. HFTs serve an investor need for continuous rebalancing. Faster trading adds opportunities to rebalance for investors. The additional marginal utility this yields is unlikely to offset the marginal cost of an HFT arms race.
9. Ultimately, I believe the economic benefits of high-frequency trading outweigh the costs. Electronic markets and HFTs arrived, and at the same time transaction costs declined for investors. This suggests that the identified economic benefits of HFTs (market making, venue competition, more trading opportunities) outweigh their economic costs (large-order predation and run games). Market quality might be further improved by exchange redesign to stimulate the economic benefits of HFTs and reduce their costs (e.g., through frequent batch auctions). Fischer Black's automated stock exchange arrived and reduced costs, but might not be in its best possible shape yet.

## FUTURE ISSUES

1. An issue that continues to concern regulators and the public at large is HFTs manipulating markets to their advantage. For example, can momentum ignition strategies (SEC 2010, p. 56) exist in equilibrium? If so, under what conditions? How might trading protocols be redesigned to curtail such practices? The same can be asked for other manipulative strategies, such as smoking or spoofing (Biais & Woolley 2011).
2. Is there systemic risk embedded in electronic trading? Do automated strategies crowd out human judgment in such a way that prices might become less informationally efficient (in the sense of less revelation of soft information, such as the quality of a new management team or the value of a patent)? In what circumstances does this occur?
3. Might machine-intermediated markets be better simply because they reduce the cost of behavioral biases present in human-intermediated markets? For example, Coates, Gurnell & Rustichini (2008) find that some human traders, propelled by high testosterone levels, are biologically predisposed to take risks.

## DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.


## ACKNOWLEDGMENTS

I benefited from many conversations on this topic over the years with experts in academia and industry, and would like to mention in particular Thierry Foucault, Terrence Hendershott, Bernard Hosman, Charles Jones, Mark Spanbroek, and my students at VU: Wenqian Huang, Sunny Li, Bart Yueshen, and Marius Zoican. I am grateful to Eric Budish, Thierry Foucault, Björn Hagströmer, Terrence Hendershott, and the editorial committee members Yacine Aït-Sahalia, Andrew Lo, and Robert Merton for invaluable comments on an earlier draft of this review. I gratefully acknowledge NWO for a VIDI grant.


## LITERATURE CITED


AFM (Neth. Auth. Financ. Mark.). 2016. *A Case Analysis of Critiques on High-Frequency Trading*. Amsterdam: Auth. Financ. Mark.

Aït-Sahalia Y, Saglam M. 2014. *High frequency traders: taking advantage of speed*. Work. Pap., Dep. Econ., Princeton Univ.

Aldrich EM, Grundfest JA, Laughlin G. 2016. *The Flash Crash: a new deconstruction*. Work. Pap., Univ. Calif., Santa Cruz

Almgren R, Chriss N. 2000. Optimal execution of portfolio transactions. *J. Risk* 3:5–39

Angel JJ, Harris LE, Spatt CS. 2015. Equity trading in the 21st century: an update. *Q. J. Finance* 5:1550002

Baldauf M, Mollner J. 2015. *High-frequency trading and market performance*. Work. Pap., Stanford Univ.

Baron M, Brogaard J, Hagströmer B, Kirilenko A. 2015. *Risk and return in high-frequency trading*. Work. Pap., Univ. Wash.

Baruch S, Glosten LR. 2016. *Strategic foundation for the tail expectation in limit order book markets*. Work. Pap., Columbia Univ.

Benos E, Brugler J, Hjalmarsson E, Zikes F. 2015. *Interactions among high-frequency traders*. Work. Pap. 523, Bank Engl., London

Benos E, Sagade S. 2016. Price discovery and the cross-section of high-frequency trading. *J. Financ. Mark.* 30:54–77

Bernales A. 2014. *Algorithmic and high frequency trading in dynamic limit order markets*. Work. Pap., Univ. Chile

Biais B, DeClerck F, Moinas S. 2015. *Who supplies liquidity, how and when?* Work. Pap., Toulouse Univ.

Biais B, Foucault T. 2014. HFT and market quality. *Bank. Mark. Invest.* 128:5–19

Biais B, Foucault T, Moinas S. 2015. Equilibrium fast trading. *J. Financ. Econ.* 73:3–36

Biais B, Glosten L, Spatt C. 2005. A survey of microfoundations, empirical results, and policy implications. *J. Financ. Mark.* 8:217–64

Biais B, Woolley P. 2011. *High frequency trading*. Work. Pap., Toulouse Univ.

Black FS. 1971a. Towards a fully automated exchange, part I. *Financ. Anal. J.* 27(4):29–34

Black FS. 1971b. Towards a fully automated exchange, part II. *Financ. Anal. J.* 27(6):24–87

Boehmer E, Fong K, Wu J. 2014. *International evidence on algorithmic trading*. Work. Pap., EDHEC Bus. Sch.

Boehmer E, Li D, Saar G. 2015. *Correlated high-frequency trading*. Work. Pap., Cornell Univ.

Bongaerts D, Kong L, Van Achter M. 2015. *Trading speed competition: Can the arms race go too far?* Work. Pap., Rotterdam Sch. Manag.

Bongaerts D, Van Achter M. 2015. *High-frequency trading and market stability*. Work. Pap., Rotterdam Sch. Manag.

Boulatov A, Bernhardt D, Larionov I. 2016. *Predatory and defensive trading in a dynamic model of optimal execution by multiple traders*. Work. Pap., High. Sch. Econ., Moscow

Breckenfelder J. 2013. *Competition between high-frequency traders, and market quality*. Work. Pap., Eur. Cent. Bank, Frankfurt, Ger.

Brogaard J, Carrion A, Moyaert T, Riordan R, Shkilko A, Sokolov K. 2015a. *High-frequency trading and extreme price movements*. Work. Pap., Univ. Wash.

Brogaard J, Garriott C. 2015. *High-frequency trading competition*. Work. Pap., Univ. Wash.

Brogaard J, Hagströmer B, Nordén L, Riordan R. 2015b. Trading fast and slow: colocation and liquidity. *Rev. Financ. Stud.* 28:3407–43

Brogaard J, Hendershott T, Hunt S, Ysusi C. 2014. High-frequency trading and the execution costs of institutional investors. *Financ. Rev.* 49:345–69

Brogaard J, Hendershott T, Riordan R. 2014. High frequency trading and price discovery. *Rev. Financ. Stud.* 27:2267–306

Brogaard J, Hendershott T, Riordan R. 2015. *Price discovery without trading: evidence from limit orders.* Work. Pap., Univ. Calif., Berkeley

Brogaard J, Hendershott T, Riordan R. 2016. High frequency trading and the 2008 short sale ban. *J. Financ. Econ.* In press

Brunnermeier MK, Pedersen LH. 2005. Predatory trading. *J. Finance* 60:1825–63

Budish E, Cramton P, Shim J. 2015. The high-frequency trading arms race: frequent batch auctions as a market design response. *Q. J. Econ.* 130:1547–621

Cappon A. 2014. The brokerage world is changing, who will survive? *Forbes*, April 16. **http://www.forbes.com/sites/advisor/2014/04/16/the-brokerage-world-is-changing-who-will-survive**

Cardella L, Hao J, Kalcheva I, Ma YY. 2014. Computerization of the equity, foreign exchange, derivatives, and fixed-income markets. *Financ. Rev.* 49:231–43

Carrion A. 2013. *Very fast money: high-frequency trading on the NASDAQ*. Work. Pap., Lehigh Univ.

Cartea A, Payne R, Penalva J, Tapia M. 2016. *Ultra-fast activity and market quality*. Work. Pap., Univ. Carlos III Madr.

Cartea A, Penalva J. 2012. Where is the value in high frequency trading? *Q. J. Finance* 2:1–46

Cespa G, Vives X. 2015. *The welfare impact of high frequency trading*. Work. Pap., City Univ. Lond.

Chaboud A, Chiquoine B, Hjalmarsson E, Vega C. 2014. Rise of the machines: algorithmic trading in the foreign exchange market. *J. Finance* 69:2045–84

Chaboud A, Hjalmarsson E, Vega C. 2015. *The need for speed: minimum quote life rules and algorithmic trading*. Work. Pap., Fed. Reserve Board, Washington, DC

Chordia T, Goyal A, Lehmann BN, Saar G. 2013. High-frequency trading. *J. Financ. Mark.* 16:637–45

Coates JM, Gurnell M, Rustichini A. 2008. Second-to-fourth digit ratio predicts success among high-frequency financial traders. *PNAS* 106:623–28

Cochrane J. 2012. Weird stuff in high frequency markets. *Grumpy Economist Blog*, Feb. 28. **http://johnhcochrane.blogspot.nl/2012/02/weird-stuff-in-high-frequency-markets.html**

de Jong F, Rindi B. 2009. *The Microstructure of Financial Markets*. Cambridge, UK: Cambridge Univ. Press

Ding S, Hanna J, Hendershott T. 2014. How slow is the NBBO? A comparison with direct exchange feeds. *Financ. Rev.* 49:313–32

Du S, Zhu H. 2014. *Welfare and optimal trading frequency in dynamic double auctions*. Work. Pap., Mass. Inst. Technol.

Easley D, López de Prado MM, O'Hara M. 2011. The microstructure of the 'Flash Crash': flow toxicity, liquidity crashes, and the probability of informed trading. *J. Portf. Manag.* 37:118–28

Easley D, López de Prado MM, O'Hara M. 2012. Flow toxicity and liquidity in a high frequency world. *Rev. Financ. Stud.* 25:1457–93

Easley D, López de Prado MM, O'Hara M. 2013. *High-Frequency Trading*. London: Risk Books

Egginton J, Van Ness BF, Van Ness RA. 2016. Quote stuffing. *Financ. Manag.* 45:583–608

Fishe R, Haynes R, Onur E. 2015. *Anticipatory traders and trading speed*. Work. Pap., Commod. Futur. Trading Comm., Washington, DC

Foucault T, Hombert J, Roşu I. 2016. News trading and speed. *J. Finance* 71:335–82

Foucault T, Kozhan R, Tham WW. 2015. *Toxic arbitrage*. Work. Pap., HEC Paris

Foucault T, Menkveld AJ. 2008. Competition for order flow and smart order routing systems. *J. Finance* 63:119–58

Foucault T, Pagano M, Röell A. 2013. *Market Liquidity: Theory, Evidence, and Policy*. Oxford, UK: Oxford Univ. Press

Foucault T, Röell A, Sandås P. 2003. Market making with costly monitoring: an analysis of the SOES controversy. *Rev. Financ. Stud.* 16:345–84
Frazzini A. 2012. *Trading costs of asset pricing anomalies*. Work. Pap. 14-05, Chic. Booth Sch. Bus., Univ. Chic.
Gai J, Yao C, Ye M. 2013. *The externalities of high-frequency trading*. Work. Pap., Univ. Ill., Urbana-Champaign
Glode V, Opp C. 2016. Asymmetric information and intermediation chains. *Am. Econ. Rev.* 106:2699–721
Glosten L, Milgrom P. 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed agents. *J. Financ. Econ.* 14:71–100
Goettler R, Parlour C, Rajan U. 2009. Informed traders in limit order markets. *J. Financ. Econ.* 93:67–87
Goldstein MA, Kumar P, Graves FC. 2014. Computerized and high-frequency trading. *Financ. Rev.* 49:177–202
Gomber P, Arndt B, Lutat M, Uhle T. 2011. *High-frequency trading*. Work. Pap., Goethe Univ.
Hagströmer B, Nordén L. 2013. The diversity of high frequency traders. *J. Financ. Mark.* 16:741–70
Hagströmer B, Nordén L, Zhang D. 2014. How aggressive are high-frequency traders? *Financ. Rev.* 49:395–419
Han J, Khapko M, Kyle AS. 2014. *Liquidity with high-frequency market making*. Work. Pap., Stockh. Sch. Econ.
Harris JH, Saad M. 2014. The sound of silence. *Financ. Rev.* 49:203–30
Hasbrouck J. 2015. *High frequency quoting: short-term volatility in bids and offers*. Work. Pap., N.Y. Univ.
Hasbrouck J, Saar G. 2013. Low-latency trading. *J. Financ. Mark.* 16:646–79
Hendershott T, Jones CM, Menkveld AJ. 2011. Does algorithmic trading improve liquidity? *J. Finance* 66:1–33
Hendershott T, Menkveld AJ. 2014. Price pressures. *J. Financ. Econ.* 114:405–23
Hendershott T, Riordan R. 2013. Algorithmic trading and information. *J. Financ. Quant. Anal.* 48:1001–24
Hirschey NH. 2016. *Do high-frequency traders anticipate buying and selling pressure?* Work. Pap., Lond. Bus. Sch.
Hoffmann P. 2014. A dynamic limit order market with fast and slow traders. *J. Financ. Econ.* 113:156–69
Hu GX, Pan J, Wang J. 2014. *Early peek advantage?* Work. Pap., Mass. Inst. Technol.
Jarnecic E, Snape M. 2014. The provision of liquidity by high-frequency participants. *Financ. Rev.* 49:371–94
Jarrow R, Protter P. 2012. A dysfunctional role of high frequency trading in electronic markets. *Int. J. Theor. Appl. Finance* 15:219–49
Johnson B. 2010. *Algorithmic Trading & DMA*. London: 4Myeloma Press
Jones CM. 2013. *What do we know about high-frequency trading?* Work. Pap., Columbia Univ.
Jovanovic B, Menkveld AJ. 2015a. *Dispersion and skewness of bid prices*. Work. Pap., Vrije Univ. Amst.
Jovanovic B, Menkveld AJ. 2015b. *Middlemen in limit-order markets*. Work. Pap., Vrije Univ. Amst.
Kirilenko AA, Kyle A, Samadi M, Tuzun T. 2014. *The Flash Crash: the impact of high frequency trading on an electronic market*. Work. Pap., Univ. Md.
Kirilenko AA, Lo AW. 2013. Moore's law versus Murphy's law: algorithmic trading and its discontents. *J. Econ. Perspect.* 27:51–72
Korajczyk RA, Murphy D. 2016. *High frequency market making to large institutional trades*. Work. Pap., Northwest. Univ.
Krugman P. 2009. Rewarding bad actors. *New York Times*, Aug. 2, p. A21
Kumar P, Seppi D. 1994. Information and index arbitrage. *J. Bus.* 69:481–509
Kyle A. 1985. Continuous auctions and insider trading. *Econometrica* 53:1315–35
Latza T, Marsh I, Payne R. 2014. *Fast aggressive trading*. Work. Pap., Cass Bus. Sch.
Laughlin G. 2014. Optical data transmissions. *Systemic Blog*, Nov. 28. **http://oklo.org/2014/11/28/optical-data-transmission/**
Laughlin G, Aguirre A, Grundfest J. 2014. Information transmission between financial markets in Chicago and New York. *Financ. Rev.* 49:283–312
Laumonier A. 2016. HFT in the banana land. *Sniper in Mahwah Blog*, Jan. 26. **https://sniperinmahwah.wordpress.com/2016/01/26/hft-in-the-banana-land/**
Lee CM, Radhakrishna B. 2000. Inferring investor behavior: evidence from TORQ data. *J. Financ. Mark.* 3:83–111
Lewis M. 2014. *Flash Boys*. New York: W.W. Norton & Company
Li S. 2015. *Long lived flow of private information, high frequency competition, and market efficiency*. Work. Pap., Secur. Exch. Comm., San Francisco

Li W. 2014. *High frequency trading with speed hierarchies*. Work. Pap., Johns Hopkins Univ.

Madhavan AN. 2012. Exchange-traded funds, market structure, and the Flash Crash. *Financ. Anal. J.* 68:20–35

Malinova K, Park A. 2016. *"Modern" market makers*. Work. Pap., Univ. Tor.

Malinova K, Park A, Riordan R. 2013. *Do retail traders suffer from high-frequency traders?* Work. Pap., Univ. Tor.

Malkiel B. 2009. High-frequency trading is a natural part of market evolution. *Financ. Times*, Dec. 14

McInish TH, Upson J, Wood RA. 2013. The Flash Crash: trading aggressiveness, liquidity supply, and the impact of intermarket sweep orders. *Financ. Rev.* 49:481–509

Menkveld AJ. 2013. High frequency trading and the new market makers. *J. Financ. Mark.* 16:712–40

Menkveld AJ. 2014. High frequency traders and market structure. *Financ. Rev.* 49:333–44

Menkveld AJ. 2016. *High-frequency trading viewed through an electronic microscope*. Work. Pap., Vrije Univ. Amst.

Menkveld AJ, Yueshen BZ. 2014. *Middlemen interaction and its effect on market quality*. Work. Pap., Vrije Univ. Amst.

Menkveld AJ, Yueshen BZ. 2015. *The Flash Crash: a cautionary tale about highly fragmented markets*. Work. Pap., Vrije Univ. Amst.

Menkveld AJ, Zoican MA. 2016. *Need for speed? Low latency trading and adverse selection*. Work. Pap., Vrije Univ. Amst.

Merton RC. 1969. Lifetime portfolio selection under uncertainty: the continuous-time case. *Rev. Econ. Stat.* 51:247–57

Moallemi C, Sağlam M. 2013. The cost of latency in high-frequency trading. *Oper. Res.* 61:1070–86

O'Hara M. 1995. *Market Microstructure Theory*. Malden, MA: Blackwell Publ.

O'Hara M. 2015. High frequency market microstructure. *J. Financ. Econ.* 116:257–70

Pagnotta E, Philippon T. 2015. *Competing on speed*. Work. Pap., N.Y. Univ.

Parlour CA, Seppi D. 2003. Liquidity-based competition for order flow. *Rev. Financ. Stud.* 16:301–43

Parlour CA, Seppi DJ. 2008. Limit order markets: a survey. In *Handbook of Financial Intermediation and Banking*, ed. A Boot, A Thakor, pp. 63–96. Amsterdam: Elsevier

Patterson S. 2012. *Dark Pools: The Rise of the Machine Traders and the Rigging of the U.S. Stock Market.* New York: Crown

Raman V, Robe M, Yadav PK. 2014. *Electronic market makers, trader anonymity and market fragility*. Work. Pap., Commod. Futur. Trading Comm., Washington, DC

Riordan R, Storkenmaier A. 2012. Latency, liquidity and price discovery. *J. Financ. Mark.* 15:416–37

Roşu I. 2016. *Fast and slow informed trading*. Work. Pap., HEC Paris

SEC (Secur. Exch. Comm.). 2010. *Concept release on equity market structure*. Concept Release 34-61358, File S7-02-10, Secur. Exch. Comm, Washington, DC

SEC (Secur. Exch. Comm.). 2014. *Equity Market Structure Literature Review, Part II: High Frequency Trading*. Washington, DC: Secur. Exch. Comm.

Stiglitz JE. 2014. *Tapping the brakes: are less active markets safer and better for the economy?* Work. Pap., Columbia Univ.

van Kervel V. 2015. Competition for order flow with fast and slow traders. *Rev. Financ. Stud.* 28:2097–127

van Kervel V, Menkveld AJ. 2015. *High-frequency trading around large institutional orders*. Work. Pap., Vrije Univ. Amst.

Vayanos D, Wang J. 2013. Market liquidity: theory and empirical evidence. In *Handbook of the Economics of Finance*, ed. G Constantinides, M Harris, R Stulz, pp. 1289–361. Amsterdam: Elsevier

Weller B. 2013. *Intermediation chains*. Work. Pap., Univ. Chic.

Weller B. 2016. *Efficient prices at any cost: Does algorithmic trading deter information acquisition?* Work. Pap., Kellogg Sch. Manag., Northwest. Univ.

Yang L, Zhu H. 2015. *Back-running: seeking and hiding fundamental information in order flows*. Work. Pap., Mass. Inst. Technol.

Yao C, Ye M. 2015. *Why trading speed matters: a tale of queue rationing under price controls*. Work. Pap., Univ. Ill., Urbana-Champaign

Yueshen BZ. 2015. *Market making uncertainty*. Work. Pap., INSEAD