ANNUAL REVIEWS

Annual Review of Economics

Implications of High-Frequency Trading for Security Markets

Oliver Linton and Soheil Mahmoodzadeh

Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, United Kingdom; email: obl20@cam.ac.uk, sm2179@cam.ac.uk

Annu. Rev. Econ. 2018. 10:237–59

First published as a Review in Advance on May 4, 2018

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

https://doi.org/10.1146/annurev-economics-063016-104407

JEL codes: G1, G2

Keywords

flash crash, high-frequency trading, liquidity, literature survey, volatility

Abstract

High-frequency trading (HFT) has grown substantially in recent years due to fast-paced technological developments and their rapid uptake, particularly in equity markets. This review investigates how HFT could evolve and, by developing a robust understanding of its effects, identifies potential risks and opportunities that HFT could present in terms of financial stability and other market outcomes such as volatility, liquidity, price efficiency, and price discovery. Despite commonly held negative perceptions, the available evidence indicates that HFT and algorithmic trading may have several beneficial effects on markets. However, these types of trading may cause instabilities in financial markets in specific circumstances. Carefully chosen regulatory measures are needed to address concerns in the shorter term. However, further work is needed to inform policies in the longer term, particularly in view of likely uncertainties and lack of data. This work will be vital in supporting evidence-based regulation in this controversial and rapidly evolving field.

1. INTRODUCTION

Are computerized trading systems delivering good outcomes for investors, speculators, hedgers, and other market participants? Computer-based trading, including by algorithmic trading (AT) and high-frequency trading (HFT), is the predominant feature in current financial markets due to technological advances and market structure developments. HFT is thought to have been responsible for as much as 75% of trading volume in the United States in 2009 (see Biais & Woolley 2011, Hendershott et al. 2011). Computers allow one to automate any trading strategy and, therefore, to do it faster, to engage in much more complicated versions of strategies that already exist, and in some cases, to do things that were simply not feasible before. They do what they are told to do and they do not complain about their bonus. They have clearly improved many back office functions. In the late 1960s, the New York Stock Exchange (NYSE) experienced an increase in trading that led to a mountain of paperwork related to the clearing and settlement process; this crisis led to the closing of the entire stock market on Wednesdays and reduced trading hours on other days in the week to allow staff to clear their desks. The level of trading volume then was less than 1% of what it is now, but computer technology is generally able to process all the current information flows in a timely fashion and with high accuracy. Many authors have argued that bid–ask spreads and the costs of transacting have decreased to both retail investors and institutional investors in the past 30 years (see Jones 2013, O'Hara 2015). These sound like improvements that we should celebrate. However, there is also a dark side. The impact of HFT on market functioning has garnered increasing scrutiny, with particular impetus stemming from events such as the May 6, 2010, Flash Crash and the October 15, 2014, bond market flash event. In the former event, nearly $1 trillion temporarily evaporated in a matter of minutes. In the latter event, the market for US Treasuries (the bedrock instrument of international finance) had its fourth-largest trading day in history, with most of the volatility concentrated in a half-hour period, with no apparent macroeconomic catalyst.

It is therefore a central issue to better understand the economic role of HFT and its impact on market functioning, which is our purpose in this review. The title of our review suggests a one-way causality from HFT to security markets, but it is more accurate to think of them as evolving together. The existence, viability, innovation, and development of HFT depend intimately on the market structure and technology. We survey the academic, industry, and government literature on computer-based trading and evaluate the evidence on the functioning of financial markets. In Section 2, we discuss the nature of HFT and AT. In Section 3, we present some critiques of HFT, while in Section 4, we discuss the role of speed and information in understanding the role of HFT and AT in financial markets, and in Section 5, we discuss the issue of profitability. In Section 6, we discuss the empirical evidence about the effects of HFT on market quality in normal times and in extreme times. In Section 7, we discuss the issue of market manipulation. In Section 8, we discuss the role of big data in HFT and AT. Section 9 concludes.

2. THE ONTOLOGY OF HIGH-FREQUENCY TRADING AND ALGORITHMIC TRADING

In this section, we try to define precisely what is meant by HFT and AT. The Securities and Exchange Commission (SEC) Concept Release of 2010 (Secur. Exch. Comm. 2010, p. 45) defines HFT thus: "(1) Professional traders acting in a proprietary capacity that generate a large number of trades on a daily basis; (2) Use of extraordinarily high speed and sophisticated programs for generating, routing, and executing orders; (3) Use of colocation services and individual data feeds offered by exchanges and others to minimize network and other latencies; (4) Very short time-frames for establishing and liquidating positions; (5) Submission of numerous orders that are cancelled shortly after submission; (6) Ending the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions overnight)." This definition was the result of many hours of committee work by top experts and has been replicated by other regulatory authorities around the world with similar results (see Foresight 2012). It emphasizes the high degree of technological sophistication and short time frames. In principle, AT includes HFT, but in practice, what is meant by the term AT, when it is placed in juxtaposition with the term HFT, is the application of computer technology for the purpose of establishing a longer-term position in the underlying securities.

According to Clarke (2017), the top 15 HFT firms worldwide in 2016 were, in order: KCG, Sun Trading, Jump Trading, Tower Research Capital, Tradebot Systems, Virtu, XR Trading, DRW Trading, GSA Capital Partners, Maven Securities, Two Sigma International, Allston Trading, IMC, Hudson River Trading, and Spot Trading. We could add Susquehanna and Citadel to this list of major firms; many banks have electronic trading desks that act similarly. This list includes regulated and even stock-market-listed firms (which are subject to substantial regulatory oversight) with hundreds of employees. In addition to these major firms, there are many smaller firms with less well-established online presences that might satisfy some or all of the SEC criteria. There is quite a bit of variation in speed within the category of traders who might be said to use HFT. Some, for example, pay for colocation, while others do not. Clearly, some of the larger firms are not acting in a purely proprietary capacity.

Perhaps a useful categorization can be based on the different strategies that HFT firms pursue. These include market making, cross-venue or cross-asset arbitrage, news-based trading, and short-term directional trading. These strategies have always existed in financial markets, but regulation and technology have affected how they are currently executed. We discuss these trading strategies in the following sections.

2.1. Market Making

According to the SEC, a market maker is a firm that stands ready to buy and sell a particular stock on a regular and continuous basis at a publicly quoted price. These firms make profits by capturing the spread between a buy order and a sell order, and the difficulty that they face is that buy and sell orders may not arrive at the same time or at the same rate. In the past, such market makers may have had a monopolistic position, knowing the order flow for their particular stock or stocks in advance of others and having little direct competition. Nowadays, this monopolistic structure has long gone. Nevertheless, exchanges like to advertise their market maker structure. On the NYSE, there are designated market makers that have obligations to maintain a fair and orderly market in their stocks, quote at the National Best Bid and Offer (NBBO) price a specified percentage of the time, and facilitate price discovery throughout the day, as well as at the open, at the close, and in periods of significant imbalances and high volatility. There are also supplemental liquidity providers that must commit to quoting a minimum quantity at the NBBO on each side of the market to meet incoming market orders in each assigned security for a certain percentage of the trading day. In return, these firms may get some advantages in terms of fees and rebates, exemption from short-sale restrictions, or even direct payments from exchanges or issuers, but they do not have exclusive or even necessarily superior access to the order flow information. Typically, these firms are large and technologically savvy, and they may participate in these arrangements for thousands of securities. In addition, there may be many participants that do not formally take part in such exchange programs for a specific security but that quote on both sides of the market in a way that is similar to official market makers, although they have the flexibility of not being required to always be in that particular market.

Unlike their textbook alter ego, current market makers use both active and passive order types, i.e., they are alternately supplying and demanding liquidity to control their inventory and risk. They need to be able to update their quotes rapidly and across all the securities in which they are active. New information about one stock suggests that quotes should be updated not only on that stock, but also on all other correlated (in practice all) securities. Market makers hope to capture the spread and do not want to offer stale quotes that give away value. Before 1999, the tick size (the minimum price increment) in the United States was one-eighth of a dollar; it is currently $0.01 for most stocks. This represents a big reduction in the lower bound of spreads that could accrue to market makers. In the more competitive environment that now operates, making markets with 1-cent spreads and no further advantage seems like a losing proposition. Assuming that market makers provide some valuable service, they need to be compensated. Consider retail foreign exchange (FX), as provided by, for example, Travelex at airports around the world. Retail FX providers typically set quotes for an entire day at a time, but their spreads are very wide, perhaps even 30% the day before an election, for example, so as to compensate for the longer resting time that they operate. They rely on the impatience and risk aversion of travelers to generate order flow and profits.

2.2. Arbitrage Activities

One popular arbitrage is between the Emini futures contract traded on the Chicago Mercantile Exchange (CME) in Chicago and the S&P 500 Index Exchange Traded Fund traded on the NYSE in New York (Budish et al. 2015). Index arbitrage exploits index tracker funds that are bound to buy and sell large volumes of securities in proportion to their changing weights in indices. If an HFT firm is able to access and process information that predicts these changes before the tracker funds do so, then it can buy up securities in advance of the trackers and sell them on to the trackers at a profit. In the FX markets, the triangular arbitrage between currencies is popular; for example, a firm can buy yen with dollars, sell yen for euros, and sell euros for dollars. If the quoted currency values are out of line, this strategy can generate profits (see Chaboud et al. 2014, Mahmoodzadeh et al. 2017).

2.3. News-Based Trading

Company news in electronic text format is available from many sources, including commercial providers like Bloomberg, public news websites, and Twitter feeds. Automated systems can identify company names, keywords, and sometimes semantics in online text, which can be used to trade news or sentiment before human traders can process it. Automated systems can also rapidly digest the implications from scheduled stock-specific and macroeconomic announcements.[1]

2.4. Direction-Based Trading

Some firms predict short-term price movements based on information embedded in market data, such as quotes, transaction prices, and volumes, as well as other sources. By buying when they predict prices will rise and then selling when they predict prices will fall, they hope to make profits from short-term movements. Momentum and contrarian strategies have been operated by many investment funds over different frequencies for a long time with mixed success (Khandani & Lo 2007).

[1] However, interpreting the subtleties of central bankers' press conferences may perhaps be beyond the current ambit of machine learning techniques.

3. SOME CRITIQUES OF HIGH-FREQUENCY TRADING

In this section, we discuss some often repeated criticisms of HFT from different quarters. There have been a number of criticisms of HFT by industry participants. The following list is based on a survey of the buy side commissioned by Foresight (2012):

1. The liquidity that HFT supplies is ephemeral. Current bid–ask spreads are narrow, but this is an illusion created by flickering quotes, i.e., the liquidity is not accessible to humans or slower traders. The high volume of trading that we see reflects so-called pass-the-parcel trading between different HFT companies (intermediation chains) rather than genuine risk transfer between final users. Furthermore, the liquidity supply of HFT evaporates rapidly during crisis times such as the Flash Crash.
2. There is a lot of implicit front-running and back-running of large institutional orders, whereby fast traders can identify incoming large orders and move ahead of them. Traders are more like ticket touts who block buy a section of the stadium before selling on to the people who actually want to go to the game.
3. There is too much messaging, i.e., order cancellations and revisions, which imposes a negative externality on other traders. The current system requires big investments in technology (smart order routers, colocation, algos for order slicing) to keep up. There is an arms race for speed (Haldane 2011).
4. The quoting and trading activities of HFT increase volatility relative to what it used to be.
5. High-frequency traders engage in market abuse and manipulation, e.g., quote stuffing, spoofing, layering, and smoking.

Many well-known economists and commentators have added their voices to the debate, including Paul Krugman (2009) in the *New York Times*:

> It's hard to imagine a better illustration [of social uselessness] than high frequency trading. The stock market is supposed to allocate capital to its most productive uses, for example by helping companies with good ideas raise money. But it's hard to see how traders who place their orders one-thirtieth of a second faster than anyone else do anything to improve that social function. . .we've become a society in which the big bucks go to bad actors, a society that lavishly rewards those that make us poorer.

Michael Lewis (2015) emphasizes that firms using HFT make money by front-running others' orders inside the stock market by being well informed about incoming orders and being faster to react and profit from them. It is convenient to use a label such as HFT to make allegations, since if one attacks any single firm or participant, then one would be subject to litigation.

We discuss further in Section 4 the questions of speed and information that are at the heart of some of these critiques. We then discuss profitability as it affects the potential scale of the damage being done by HFT.

4. SPEED AND INFORMATION

Since speed is the subject of much criticism, we review its role in financial markets. In fact, speed has always mattered to investors. A well known example of this is the legend of Nathan Mayer Rothschild profiting from news of the British and Prussian victory at Waterloo. He had news about the outcome of the battle ahead of the British government (the fantasy version has it that he received the information by carrier pigeons; other versions say he received it by personal couriers). There are also two versions of how he made money. In the first, he simply bought bonds on news of victory, while in the second version of events, he spoofed the market by publicly selling bonds

Table 1 System latency of the London Stock Exchange matching engine

System	Implementation	Latency[2] 10^{-6}
SETS	Before 2000	600,000
SETS1	November 2001	250,000
SETS2	January 2003	100,000
SETS3	October 2005	55,000
TradElect	June 18, 2007	15,000
TradElect 2	October 31, 2007	11,000
TradElect 3	September 1, 2008	6,000
TradElect 4	May 2, 2009	5,000
TradElect 4.1	July 20, 2009	3,700
TradElect 5	March 20, 2010	3,000
Millenium	February 14, 2011	113

himself while having his agents buy them and keep on buying them as the prices dropped. In any case, he made lots of money from having the key information 24 hours before other participants. He was the high-frequency trader of his day.

We have benefitted, in the past 50 years, from remarkable technological improvements. In 1965, Gordon Moore predicted that the number of transistors in a dense integrated circuit would double approximately every 2 years, and this prediction has been fairly accurate since then (although it has been claimed that it will cease to operate from 2025). This development has led to a rapid increase in computing power and speed and improved all of the applications of technology. **Table 1** illustrates the evolution of speed in the past 20 years as measured by the system latency of the London Stock Exchange's matching engine. This roughly follows Moore's law, with an approximate halfing of latency every 2–3 years. The result of this is that the trading system itself can handle many more messages and deliver execution very much faster than it could prior to 2000 and, of course, much faster than any human-intermediated system, which were common prior to the 1980s. Of course, traders have developed systems and approaches to take advantage of this technological improvement, just as academics have also improved their productivity by making use of faster computers and better software.

This shows that the trading infrastructure has become faster, which means that the costs of delivering a given speed has declined, as well. HFT is a phenomenon that has arisen throughout this technological development, as all trading-related activity has sped up. Since 2011, there has been a trend to use microwaves to transmit data across key connections, such as that between New York City and Chicago. Microwaves traveling in air suffer a less than 1% speed reduction compared to light traveling in a vacuum, whereas with conventional fiber optics, light travels over 30% slower. However, microwaves are more fragile, for example, in storms (Shkilko & Sokolov 2016). As everything speeds up, the physical distance between locations starts to matter much more. Knowing the latency delay factor is important in determining which of two messages from different exchanges was truly issued first. In this calculation, random variation in latency and clock accuracy can also be a big factor. This makes perfect calculations, as well as guaranteeing that one is first, impossible, although one may obtain a systematic advantage by having the best technology (Kirilenko & Lamacie 2015). Academics and regulators working with post-trade data face major problems in scrutinizing markets. For this analysis, one also needs to know that the clocks on each exchange are synchronized so as to work out which message came first, or, since they are not, in fact, so perfectly synchronized, one needs to know what is the precise time delay between the clocks on the exchanges.

How do we measure the benefits of speed in electronic markets? Posting limit orders on the exchange gives options to trade to other traders, since they have the right but not the obligation to execute against you (Copeland & Galai 1983). The Black and Scholes call-option price can be used to value the option. In a standard notation, the price of the call option C in terms of the underlying price S is

$$C(S, X, \tau, r_f, \sigma) = S \cdot \Phi(d_+) - X \cdot e^{-r_f \cdot \tau} \cdot \Phi(d_-),$$

$$d_\pm = \frac{\log \frac{S}{X} + \left(r_f \pm \frac{\sigma^2}{2}\right) \cdot \tau}{\sigma \cdot \sqrt{\tau}},$$

where Φ is the standard normal cumulative distribution function, while r_f is the interest rate, τ is the time to maturity, X is the strike price, and σ is volatility. A competitive limit order can be considered to be at the money, i.e., $S = X$; furthermore, intraday interest rates r_f are zero. Letting $\tau \to 0$, we obtain the approximation

$$\frac{C(S, X, \tau, r_f, \sigma)}{S} = \frac{1}{\sqrt{2\pi}} \sigma \sqrt{\tau} + O(\tau^{3/2}). \qquad 1.$$

This says that there is a positive, albeit small, value in a single order that sits for a small time (if there are many orders, then the total value being given away can be large). Also, Equation 1 says that the value being given away increases with volatility measured by the parameter σ, so that, in times of market stress, the value per unit time increases. The posters of limit orders should be compensated for their service to the market by, for example, the bid–ask spread. The faster the limit order poster is at updating their quotes, the less compensation they would require for posting them in the first place. Note that the value in Equation 1 scales with the square root of time: millisecond to microsecond value shrinks by 1/30, not by 1/1,000.

The classical Glosten & Milgrom (1985) microstructure model of a dealer market explains the bid–ask spread in terms of adverse selection. The dealers are uninformed and post limit orders for informed and uninformed traders who arrive randomly. In this model, on the one hand, if the dealers can update quotes in response to new information faster than the informed traders can act, then they will be able to set narrower spreads than if they are slow and keep stale quotes on the table. On the other hand, if the informed traders are faster than the dealers, then the dealers will have to set wider spreads to protect themselves. In the classical dealer market, the dealers alone know the order flow, and they alone set prices; investors have to take them or leave them. Dealers earn a profit from providing the service of immediacy. In the new system, that advantage has been eliminated, and without some advantage, the human and physical capital that would have gone to the market making activity would find better employment elsewhere. Speed of action and superior information about order flow from data feeds are the advantages that the new market makers seek (Menkveld 2013).

The classic inventory models, e.g., that of Ho & Stoll (1981, 1983), explain the spread as the cost of providing immediacy to impatient investors. In these models, the spread varies positively with the degree of the monopoly power of the dealer, the volatility of the asset price, the trade size, and the horizon of the dealer. This class of models predicts that competition between dealers would lead to small spreads. It also predicts that small-sized orders would require small spreads, whereas larger orders would face wider spreads, ceteris paribus. Also, if the dealer has a short horizon, then spreads should be narrow. All of these model predictions suggest that dealers who can quote and cancel faster will benefit the market. Aït-Sahalia & Saglam (2013) extend this literature to analyze the consequences for liquidity provision of competing market makers operating at high frequency. They find that competition increases overall liquidity and deters the fast market makers' use of order flow signals. They show that the market makers provide more liquidity as they get faster but shy away as volatility increases.

A lot of recent work has extended microstructure models to give a special role to high-frequency traders. In some cases, the presence of these traders delivers positive outcomes, while in others, they do damage to market quality metrics (Aït-Sahalia & Saglam 2013, Biais et al. 2015, Cartea & Penalva 2012, Foucault et al. 2016b, Jarrow & Protter 2011, Jovanovic & Menkveld 2011, Pagnotta & Philippon 2015). These new models allow firms using HFT to affect the functioning of the trading system by being better informed agents (i.e., subscribing to news feeds), leading to a positive effect; being faster acting but uninformed, leading to a negative effect; preying on large informed traders, leading to a negative effect; preying on large uninformed traders, leading to a positive effect; running games in the context of multiple markets, leading to a negative effect; or using mixed strategy over prices, leading to endogenous quote flickering, a negative effect.

These are all only scenarios that could happen; there is no dominant agreed-on model yet. By their nature, economic models have to be simplifications, and in this setting, the types of simplifications that must be made are somewhat extreme and fragile in the sense that small changes in modeling assumptions can predict quite different outcomes. Many models have no explicit time scale (for example, there are two or three periods, or the evolution of time is discrete and exogenous), when time is the key and often endogenous feature in this case. The interpretation of some models is not unique, since they present some traders with an advantage that could represent speed or simply being smarter.[2] Models often involve a simplified strategy space for traders that captures some key feature of interest. A common feature of many models is that traders are either informed or uninformed (about the true value of the security), and in many cases, this is the key feature that differentiates participants. This is quite a drastic simplification, and it is not clear that it is justified as the main driver of outcomes. In FX markets, for example, where there are no retail traders, it is not so clear whether this binary classification is useful. Kirilenko et al. (2017) record more than 15,000 trading accounts for the Emini futures contract around the time of the Flash Crash. For their empirical analysis, they collapse these traders into six broad categories based on their observed trading styles (in terms of their attitude to inventory and holding periods) over the 4 days including the flash crash. In practice, the HFT category is consistent with a number of different trading strategies, such as market making, arbitrage, and short-term directional trading, which can have quite different effects on markets.

5. PROFITABILITY

One key question is whether HFT activity is a significant and pernicious factor in the market as a whole. The TABB Group (2012) estimates that the HFT sector earned $20 billion in profits in 2008, which raises some concerns about the trading practices of these firms (see also Budish et al. 2015). Although firms using HFT make a small profit per trade on average, since they make many trades, they may extract a large amount of rent from the intermediation sector, which would negatively impact end users. Kearns et al. (2010) investigate the potential profitability of HFT. They argue that the negative consequences of HFT seem to come primarily from traders' use of aggressive market orders. Kearns et al. use the so-called omniscient trader methodology to try to estimate an upper bound on the potential profits being made by HFT through this strategy. They reconstruct the entire limit order book of Nasdaq in 2008 for 19 large firms (this was a major computational undertaking). The omniscient trader who can see the entire evolution of

[2] Perhaps we should be worrying about high-intelligence traders, especially in the era of artificial intelligence.

the market uses market orders with variable direction and quantity and holds for the time period $\in \{0.01, \ldots, 10\}$. The calculation is repeated every 10 milliseconds. Kearns et al. extrapolate to the entire equity universe by regression methodology. They find that the potential profits achievable for this strategy are quite small, quite a bit less than $2 billion per year over the equity space for this strategy. Others have argued that competition has further reduced profits in this space. Chordia et al. (2011) argue that HFT profits based on trading quickly after macroeconomic announcements have declined in recent years, consistent with more competition between traders using HFT, or low-latency traders, as they call them.

In fact, one can obtain further information on the actual profits earned by HFT. For example, KCG and Virtu are both regulated and listed on the NYSE and have to file regular accounts. Their stock prices do not imply extreme profitability, and their annual reports further show that gross revenues and executive compensation are also not as high as implied by the Tabb Group's estimates. However, some stock exchanges have made a lot of money since 2007 (in many cases by selling data and technology), judging by their stock prices and annual reports.

To summarize, HFT is believed by some to allow creaming off of huge profits from innocent asset managers and retail investors. In fact, profitability of the sector has fallen through increased competition. There are economic arguments for why HFT can cause bad outcomes for the market as a whole, and there are economic arguments for why HFT can cause good outcomes. The question of the value of HFT is ultimately an empirical one, and we turn to this in the next section.

6. EMPIRICAL EVIDENCE ABOUT MARKET QUALITY

In this section, we review the empirical evidence about the effects of HFT on the functioning of markets, i.e., so-called market quality.

Institutional networks introduced the first automated trading system in 1969, and in 1977, the Green screen, which displayed quotes from the NYSE, was introduced. Glosten (1994) asks whether the electronic limit order book (ELOB) was inevitable, and by 2000, it was a substantial part of the landscape. It is commonly assumed that the era of HFT began around 2005, although given the difficulties in defining HFT, this is not so firm a dating.[3]

To determine the effect of HFT on outcomes, one could compare the market outcomes before and after 2005. However, this would surely be oversimplified because the global financial crisis of 2007–2010 affected many financial and economic outcomes over the same time period and in a much bigger way than did HFT.[4]

Financial markets have changed in many ways to reflect technological advances and regulatory changes. The encouragement of competition between trading venues brought about by the Regulation National Market System (Reg NMS) in the United States and Directive concernant les services d'investissement/MiFID in Europe led to a more diverse financial ecosystem, which led to improvements in market quality (Gresse 2011, O'Hara & Ye 2011). There are currently more trading venues than there were 20 years ago, and the diversity of trading venue types has also increased. Best execution policy in the United States forces some integration of the trading venues by imposing the law of one price for a small quantity. It fosters competition so that, for example, the NYSE cannot simply trade through a better quote placed elsewhere. This integration is accomplished by smart order routing technology that links markets together (Foucault & Menkveld

[3] According to Google Trends, the term HFT became frequently searched around July 2009; searches continued to increase after that and reached their peak in 2014.

[4] If HFT has a role to play in the large swings in market conditions, it is relatively small and insignificant in comparison with the huge negative effects of the banking and sovereign debt crises that happened during the financial crisis.

2008). There has been a substantial development of algorithmic software to effectuate a variety of trading strategies. These algorithms are given names such as Stealth, Iceberg, Dagger, Guerrilla, Sniper, and Sniffer. They are routinely bought or rented by a range of participants, along with the technology to implement them. There has also been a development of electronic dark pools. These are alternative trading systems that are private in nature—and thus do not interact with public order flow—and seek to provide undisplayed liquidity to large blocks of securities. In dark pools, trading takes place anonymously, prices and quantities are not displayed as in the lit venues of standard exchanges, and execution prices are usually set at the midpoint of the best bid and offer from some lit venue or venues. Some authors have questioned whether dark pools degrade the overall market quality by impeding price discovery (Degryse et al. 2015). More recently, there has been concern as to whether the transaction prices achieved in dark pools are accurately pegged to current best quoted prices (Aquilina et al. 2017).

There have been many other changes in markets before and after 2005. The decimalization process in the US markets, which saw the tick size reduced from 12.5 cents to 1 cent for large stocks around 2000, is a significant factor in the decline of bid–ask spreads, since in the old regime, no matter how much competition there was, the spread could not go below 12.5 cents (Aït-Sahalia & Yu 2009, Bessembinder 2003).[5] The demutualization process of the exchanges themselves, which are now listed companies rather than companies owned by their users, has lead them to aggressively pursue pricing strategies to encourage trading on their marketplace (Malinova & Park 2011) and to improve their technology. The introduction of competition between exchanges in Europe in 2007 after MiFID and the intensification of competition between exchanges in the Untied States following the introduction of Reg NMS are potential factors in explaining the improvement in market quality metrics (Storkenmaier & Wagener 2011).[6] The growth of the economies and stock markets of Brazil, Russia, India, and China, however, has led to more sensitivity of prices to news from around the world.

6.1. Measurement of High-Frequency Trading and Market Quality

We discuss above the difficulty in defining HFT. In the empirical literature, there are generally two ways to measure HFT activities. There are proxies (like message traffic) that are related to the intensity of trading activity. Large volumes of order placements and cancellations indicate at least the involvement of computers. Examples using US data can be found in the work of Brogaard et al. (2014) and Hendershott et al. (2011). One issue with this approach is that it can be a poor guide to the presence or absence of HFT.

A key issue in identifying the consequences of HFT for market quality is endogeneity. That is, property x may be both a cause and a consequence of HFT activity. The question is whether HFT causes more volatility or volatility causes higher activity by traders using HFT (this can be the case since volatility offers profitable trading opportunities to some users of HFT). The econometric methods available to identify the direction of causation, or rather, to control for endogeneity, are as follows. In the first approach, difference in differences may be taken, that is, the difference between treated and untreated outcomes before and after the treatment are compared. This approach eliminates common time-specific and firm-specific effects that may contaminate

[5] In the Electronic Broking Services (EBS) FX market after 2011, the tick size was reduced by a factor of 10 from 0.0001 (pips) to 0.00001 (decimal pips). In Bitcoin, the minimum price increment is 10^{-8}.

[6] MiFID2 will be implemented in January 2018. It will bring more trading onto electronic venues (from over the counter) and impose larger minimum size orders for dark pools, with the aim of improving transparency.

our view of the effects of HFT. However, it is sometimes difficult to find a proper control group and to justify the parallel trends assumption (Bertrand et al. 2004). A second approach is to use an instrumental variable, that is, a variable that is related to the input variable (for example, HFT activity) but unrelated to the output variable (for example, market quality) except directly through the input variable (HFT). The main problem with this approach is finding credible instruments, that is, instruments that are not correlated with the error term. Popular instrumental variables include latency upgrade events, such as those listed in **Table 1**, or the time at which an exchange first adopted automation.

6.2. Liquidity

Liquidity is a fundamental property of a well-functioning market, and lack of liquidity is generally at the heart of many financial crises and disasters. Is HFT associated with a decrease or increase in liquidity during regular market conditions? There have been substantial and well-documented changes to some important features of stock market trades and quotes over the past 12 years: The average size of transactions has decreased, the number of transactions has increased, the number of quotes has increased, the number of quotes per transaction has increased, and the average holding period of stocks has decreased (although perhaps not to the 11 seconds of Internet rumor) (Chordia et al. 2011). However, these metrics are not generally interpreted by themselves as measures of liquidity or market quality. Common ways of measuring liquidity include bid–ask spreads, effective spreads, realized spreads, depth and weighted depth, transaction volume, and Amihud illiquidity (see Goyenko et al. 2009).

Several studies try to identify computerized trading and its consequences for the order book and transactions. Hendershott et al. (2011) use the automation of the NYSE quote dissemination as an implicit experiment to measure the causal effect of AT on liquidity. In 2003, the NYSE began to phase in the auto-quote system, which empowered computerized trading, initially for six large active stocks and then slowly, over the next 5 months, for all stocks on the NYSE. Hendershott et al. find that this change narrowed spreads, which they interpreted as increasing AT, improving liquidity, and reducing adverse selection. The evidence was strongest for large stocks. Another study by Chaboud et al. (2014) also reports results on liquidity in the EBS exchange rate market. They find that, even though some algorithmic traders appear to restrict their activity in the minute following macroeconomic data releases, they increase their supply of liquidity over the hour following each release.

Hasbrouck & Saar (2013) investigate order book data from Nasdaq during the trading months of October 2007 and June 2008. Looking at 500 of the largest firms, they construct a measure of HFT activity by identifying strategic runs, which are linked submissions, cancellations, and executions. These are likely to be parts of a dynamic strategy adopted by such traders using HFT. Their conclusion is that increased low-latency activity improves traditional market quality measures such as spreads and displayed depth in the limit order book, as well as reducing short-term volatility.

Brogaard (2010) also investigate the impact of HFT on market quality in US markets. High-frequency traders were found to participate in 77% of all trades and tended to engage in a price-reversal strategy. There was no evidence to suggest that high-frequency traders were withdrawing from markets in bad times or engaging in abnormal front-running of large non-HFT trades. High-frequency traders demanded liquidity for 50.4% and supplied liquidity for 51.4% of all trades. They also provided the best quotes approximately 50% of the time.

In Europe, Menkveld (2013) study in some detail the entry of a new high-frequency trader into trading on Dutch stocks at Euronext and a new market, Chi-X, in 2007 and 2008. He shows that the inventory of the high-frequency trader ends the day close to zero but varies throughout the day, which is consistent with the SEC definition of HFT. All of the trader's earnings arose from passive orders (liquidity supply). He also finds that the bid–ask spreads were reduced by approximately 30% within a year when compared with Belgian stocks that were not traded by the HFT entrant. Brogaard & Garriott (2017) show similarly improved spread metrics on the Alpha exchange in Canada after the arrival of 11 HFT firms.

There are also studies reporting trends in liquidity without specifically linking them to AT or HFT. Castura et al. (2010) investigate trends in bid–ask spreads on the Russell 1000 and 2000 stocks over the period 2006–2010. They show that bid–ask spreads declined over this period and that available liquidity (defined as the value available to buy and sell at the inside bid and ask) improved over time. Angel et al. (2015) show a slow decrease in the average spread for S&P 500 stocks over the period 2003–2010 (subject to some short-term up-side fluctuations in 2007–2008). They also find that depth increased slowly over the relevant period. The evidence also shows that both the number of quotes per minute and the cancellation-to-execution ratio increased, and market order execution speed increased considerably.

Friedrich & Payne (2015) compare the operation of HFT in equities and FX. They find that penetration of algorithmic, dynamic agency flow (i.e., best execution of trades on behalf of clients) on multilateral order books in FX is small relative to equities, perhaps because FX is more liquid and therefore orders do not need to be broken up. They report no trend in volume (the traded value) of FTSE 100 stocks traded between 2006 and 2011 but find that bid–ask spreads decreased, while depth increased. The number of trades, however, increased more than five times over this period, implying that the average trade size is now only 20% of its former level. For small UK stocks, there are different results. First, the average trade size has not changed as much over the period 2006–2011, which suggests that HFT is not so actively involved in UK trading. Second, there has been little improvement in the liquidity of small cap stocks.

6.3. Transaction Costs

Trading with computers is cheaper than trading with humans, so transaction costs have fallen steadily in recent years as a result of the automation of markets. Jones (2002) reports the average relative one-way costs paid for trading Dow Jones stocks between 1935 and 2000. He finds that the total cost of trading has fallen dramatically in the period 1975–2000. Angel et al. (2015) show that average retail commissions in the United States decreased between 2003 and 2010, a period more relevant for inferring the effects of computer trading. They also make a cross-country comparison of trading costs at the end of 2009. According to this study, the US large cap stocks are the cheapest in the world to trade, with an approximately 40-basis-point cost. The US marketplace is the most impacted by technology and HFT.

Menkveld (2013) argues that new entry, often designed to accommodate HFT, has profound effects on transaction costs. For example, the entry of Chi-X into the market for Dutch index stocks had an immediate and substantial effect on trading fees for investors, first through the lower fees that Chi-X charged and then through the consequent reduction in fees that Euronext offered. The strongest effect, however was a reduction in clearing fees. A new clearing house, EMCF, entered in 2017, and this triggered a price war that ended with a 50% reduction in clearing fees. This reduction seems to have been replicated across European exchanges, to the benefit of investors.

The interests of institutional investors are of great importance. Brogaard et al. (2012) examines the direct effects of HFT on the execution costs of long-term investors. The authors use a new UK data set obtained from the detailed transaction reports of the Financial Services Authority over the period 2007–2011 to provide a better measurement of HFT activity. They combine this data set with Ancerno data on institutional investors' trading costs. To test whether HFT has impacted the execution costs of institutional traders, the authors conduct a series of event studies around changes in network speeds on the London Stock Exchange to isolate sudden increases in HFT activity. This study finds that the increases in HFT activity have no measurable effect on institutional execution costs. Of course, additional studies linking HFT and institutional trading costs in other market settings would be helpful in determining the generality of this finding.

6.4. Price Discovery and Efficiency

The usual method of measuring the degree of market inefficiency is by using the predictability of prices based on past price information alone. In practice, widely used measures such as variance ratios and autocorrelation coefficients estimate the predictability of prices based on linear rules. Hendershott (2012) describes the meaning of price efficiency in the context of high-speed markets and presents arguments for why HFT may improve market efficiency by enabling price discovery through information dissemination. Brogaard et al. (2014) find that high-frequency traders play a positive role in price efficiency by trading in the direction of permanent price changes and in the opposite direction of transitory pricing errors on average days and the days of highest volatility. Negative effects on efficiency can arise if high-frequency traders pursue market manipulation strategies (see below). However, it is clear that price efficiency–reducing strategies, such as manipulative directional strategies, are more difficult to implement effectively if there are many firms following the same strategies. Thus, the more competitive is the HFT industry, the more efficient will be the markets in which traders operate.

A variety of evidence suggests that price efficiency has generally improved with the growth of computer-based trading. Castura et al. (2010) investigate trends in market efficiency in Russell 1000 and 2000 stocks traded on the NYSE and Nasdaq over the period from January 1, 2006 to December 31, 2009. They compare the variances of stock returns computed at 1 second, 10 seconds, 1 minute, and 10 minutes. Prior to the automation of the NYSE in 2006–2007, the NYSE had much slower trading than Nasdaq, which meant it was less attractive to high-frequency traders. As the automation proceeded, penetration by HFT increased on the NYSE. Based on evidence from intraday variance ratios, Castura et al. argue that markets became more efficient in the presence of and with increasing penetration by HFT. In summary, the preponderance of evidence suggests that HFT has not harmed, and may have improved, price efficiency.

6.5. Volatility and Stability

There is concern that some HFT systems, like other novel trading systems in the past, could be making a steady stream of small profits but at the risk of causing very big losses if (or when) things go wrong, i.e., picking up pennies before steamrollers, as the saying goes. Even if each individual HFT system is considered to be stable, it is well known that groups of stable systems can, in principal, interact in highly unstable ways. Price volatility is a fundamental measure that is useful in characterizing financial stability, since wildly volatile prices are a possible indicator of instabilities in the market and may discourage stock market participation.

Many authors have argued that the introduction of computerized trading and the increased prevalence of HFT strategies in the period after 2005 has lead to an increase in volatility (see Benos & Sagade 2012, Boehmer et al. 2012, Caivano 2015, Zhang 2010). One of the key issues that needs to be addressed in making such a comparison is the time frame under consideration. For example, retail FX providers such as Travelex keep their midquote constant throughout a trading day, and common measures of intraday volatility calculated from this price would be zero, whereas the spot or futures FX markets would reveal nontrivial and time-varying intraday volatility. However, day-to-day variability in Travelex midquotes would essentially track the volatility on the spot market. The time frame for comparison is critical.

How can one test this hypothesis? Several studies have investigated this question with natural experiments methodology (e.g., Brogaard et al. 2014, Hendershott & Riordan 2013), but the conclusions that one can draw from such work are event specific. One implication of this hypothesis is that, ceteris paribus, the ratio of intraday to overnight volatility should have increased during this period because trading is not taking place during the market close period. Linton & Wu (2016) show that, for large stocks, the reverse has happened, i.e., the ratio of overnight to intraday volatility has increased over the period 2001–2016. This finding seems to be hard to reconcile with the view that trading has increased volatility. Hasbrouck (2018) compares the volatility of quoted prices over the 2001–2011 period. At subsecond horizons, bids and offers in US equity markets are more volatile than what would be implied by long-term fundamentals. He suggests that traders' random latencies interact with quote volatility to generate execution price risk and relative latency costs and that this volatility is more likely to arise from recurrent cycles of undercutting, rather than mixed strategies of limit order placement. He also shows that this quote volatility does not display a strong trend despite the high growth in quote traffic. Overall, the evidence does not support the view that HFT has increased volatility in normal times.

We now turn to the discussion of extreme events or flash crashes. Flash crashes are short and relatively deep price movements that are not apparently driven by fundamentals, or rather, they are movements in prices that are in excess of what would be warranted based on fundamentals, at least according to hindsight. Unlike in the case of some other market crashes (e.g., those in 1929 and 1987), one may not easily identify a prior period where the market was dominated by bubbles. Also, in many cases, flash crashes are not contagioned globally, unlike, say, the 1929 and 1987 crashes, which were worldwide phenomena. Some argue that, in certain specific circumstances, self-reinforcing nonlinear feedback loops (the effect of a small change looping back on itself and triggering a bigger change, which again loops back, and so on) within well-intentioned management and control processes can amplify internal risks and lead to undesired interactions and outcomes (Danielsson & Shin 2003). These feedback loops can involve risk-management systems and can be driven by changes in market volume or volatility, by market news, and by delays in distributing reference data. HFT has the potential to lead to a qualitatively different and more obviously nonlinear financial system in which crises and critical events are more likely to occur in the first place, even in the absence of larger or more frequent external fundamental shocks. In the Glosten & Milgrom (1985) class of models, flash crashes may be caused by increased toxic order flow from informed agents or a misinterpretation of a large temporary directional order flow as being permanent.

We first consider the Flash Crash in the US stock market on May 6, 2010. In **Figure 1**, we show the trajectory of the Emini futures price in the hour containing the peak declines and rises, along with the change in consecutive transaction prices in terms of ticks. This shows the rapidity of the price changes and the incredible volatility that was present during the crash. Even a half hour before the crash, most price changes between consecutive transactions took place within one tick, but during the most intense period, there were price changes of up to 40 ticks in both directions.

The report by the SEC and the Commodity Futures Trading Commission (Commod. Futures Trading Comm. & Secur. Exch. Comm. 2010) on the Flash Crash suggests some explanations for its initiation and promulgation. The report argues that the starting point was a large parent sell order by Waddell & Reed for 75,000 Emini contracts that was divided into price-insensitive market orders and fed into the market at a rate proportional to the transaction volume that had occurred in the most recent period. This led to a dynamic interaction between different types of

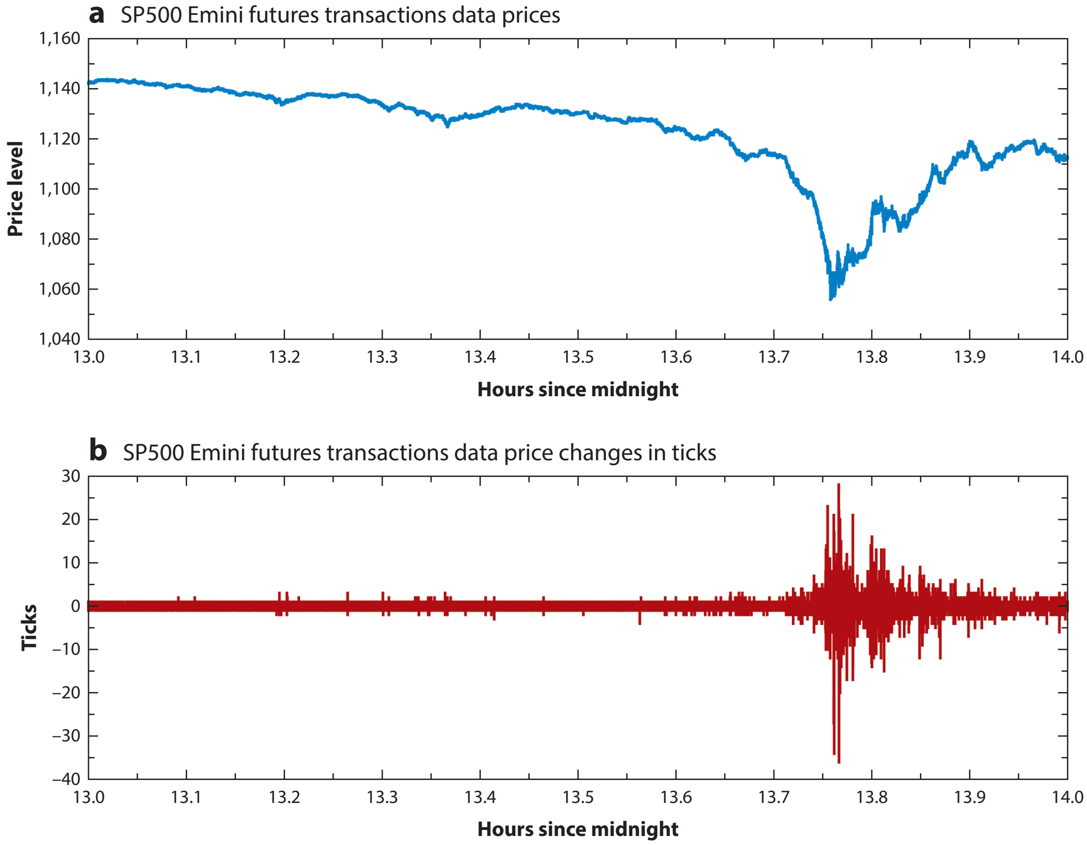

Figure 1

(*a*) Price level of the Emini near-term futures contract during the Flash Crash along with (*b*) changes in prices between consecutive transactions measured in ticks.

traders as they tried to absorb the large volume by trading among themselves, which, in turn, raised the transaction volume, which led to more selling by the Waddell & Reed algo. At some point, the loop broke, and high-frequency traders withdrew liquidity. Kirilenko et al. (2017) conclude that high-frequency traders did not trigger the Flash Crash but that their responses to the unusually large selling pressure on that day exacerbated market volatility. Easley et al. (2011) argue that historically high levels of order toxicity forced market makers to withdraw during the Flash Crash.[7] Others have disputed some parts of this narrative (Andersen & Bondarenko 2014, Hunsader 2010, Madhavan 2011, Menkveld & Yueshen 2017).

Some other recent technological and market disasters include the Facebook initial public offering (IPO) (#Faceplant) on May 18, 2012, when trading was delayed for a half hour; the BATS IPO (the stock opened at $15.50 but traded down to 1 cent in 1.4 seconds); and the Google (#PendingLarry) on on line mistaken early earnings announcement in 2012 (the price went down 10% in 8 minutes). During the so-called Knightmare on Wall Street on August 1, 2012, Knight

[7] Order flow is considered toxic when it adversely selects market makers who are unaware that they are providing liquidity at their own loss.

Capital, a market maker and HFT firm, listed on the NYSE. A trading error caused widespread disruption on the NYSE, and the firm lost $450 million in a few minutes (see Nanex 2012). Subsequently, the firm was bought out by another HFT firm. During the so-called Hash Crash on April 23, 2013, the Reuters Twitter account was hacked and a story tweeted about a bomb at the White House, which resulted in a rapid drop in the Dow Jones index and subsequent rapid recovery when the hack was uncovered. The Treasury Flash Event (where prices went up rapidly and yields went down equally rapidly) of October 15, 2014, was a very big event, and it has proven difficult to explain the magnitude of the short-term price changes.

FX markets have some different features from equity markets. They have a large OTC component, where the identity of the transacting parties is common knowledge (unlike in the electronic order book, where price and quantity are displayed but identity is hidden), and they are lightly regulated. There have been a number of recent rule changes on EBS, one of the largest electronic platforms, aiming to limit and mitigate the actions of anonymous HFT participants. EBS introduced a minimum quote life in 2009, mandating that quotes must remain tradable for at least a minimum amount of time (e.g., 500 milliseconds) to give most participants the opportunity to trade on them. A latency floor was introduced in 2014 in which orders arriving with a period (which is itself random in length and start time) are kept within one or more batches and their priority within the batch is randomised (ParFX and Reuters also have versions of this system). Nevertheless, flash crashes have occurred in FX markets. One major event was the Swiss franc depegging event of January 15, 2015, in which the Swiss authorities removed their large limit order on EBS that was supporting the desired price level. After a short while, this resulted in a rapid increase in the value of the Swiss franc from 1.20 to 0.80 and a subsequent recovery to 1.05. The downward trajectory was chaotic, and there was gapping, meaning that many price points between between the high and low points were not visited. Although the primary cause of the crash is obvious and reflects fundamental information, the way in which it transpired is worrisome.

The Sterling Flash Crash on October 7, 2016 is a more recent example. The GBP/USD currency, the third most liquid currency pair in the world, dropped by 9.66%, from 1.2601 to 1.1491, within 40 seconds. Most of this movement was reversed within the 10 minutes that followed. The Bank for International Settlements provides a report (Bank Int. Settl. 2017) on the Sterling Flash Crash. Rather than pointing to any single driver, it finds the movement in the currency pair to have resulted from a "confluence of factors" (Bank Int. Settl. 2017, p. 1). These included larger-than-normal trading (predominantly selling) volumes at a typically illiquid part of the trading day, as well as demand to sell sterling to hedge options positions and execute client orders in response to the initial fall in the exchange rate. The report also notes the potential amplifying role played by trading halts in GBP/USD futures contracts on the CME futures exchange. This may have created larger price pressure on other platforms and increased the price impact of trades in the spot market because many trading systems rely on the linkage between the spot and futures markets.

Kyle & Obizhaeva (2016b) propose a comprehensive model of trading in markets as transfers of risk in business time. Kyle & Obizhaeva (2016a) apply their findings to explain crashes of the market in October 1929, October 1987, and January 2008, as well as the flash crash of May 2010. They define the concept of bets as a parent sell or buy order that can be decomposed into many small trades. The bet does not have to be placed by a single person but can be composed of multiple orders acting on the same impulse or information. An example of this is when a price decline forces a liquidation of positions made on margin in a given security. The advantage of this approach is that it does not suffer from biases caused by analyzing all the decomposed small trades (which individually have small impact) separately. Tobek et al. (2017) argue that the first part of the Sterling Flash Crash was broadly consistent with the Kyle & Obizhaeva (2016a) impact model, given the large directional order flow, but that subsequent price developments went beyond what would be countenanced by the impact model.

There has been an increase in the deployment and variety of circuit breakers and other market controls that limit price movements or halt trading on exchanges. Their purpose is to reduce the risk of a market collapse induced by a sequence of cascading trades. Circuit breakers can take many forms. There have been three types of circuit breakers in the US equity markets: the market-wide circuit breaker, the single-stock circuit breaker, and the limit up–limit down trading halt. The market-wide circuit breaker shuts down all trading when triggered by a large movement in the stock price index. The single-stock circuit breaker shuts down trading (or switches trading to an auction mechanism) when there is a large movement in the individual stock price. The limit up–limit down mechanism prohibits trade outside upper and lower bounds; however, trade of the stock can continue within the limits. Some authors have found positive effects of circuit breakers on market quality (Brugler & Linton 2017).

Circuit breakers are no panacea. Price discovery is a natural feature of markets, and bad news can induce (sometimes large) price drops to new efficient values. Halting markets can interfere with this natural process and may simply postpone the inevitable. An empirically documented effect of circuit breakers is the so-called magnet or gravitational effect, whereby traders rush to carry out trades when a halt becomes imminent, accelerating the price change process and forcing trading to be halted sooner or moving a price more than it otherwise would have moved (Arak & Cook 1997, Subrahmanyam 1994). Some have argued that, during the Sterling Flash Crash, the implementation of trading halts in the CME futures market contributed to the collapse of the spot market. In January 2016, China suspended its recently implemented circuit breaker system after it experienced 2 days in close succession when trading in Chinese stocks had been halted because of a plunge in prices. The China Securities Regulatory Commission has used the halts and other measures to control downward pressure amid volatility. However, some observers felt the system as designed could have increased investor jitters about the health of markets.

7. MARKET MANIPULATION

Markham (2014) describes the long history of market manipulation, including squeezes, pump and dump schemes, and insider trading, in the era before and including electronic trading. The following types of market manipulation are of particular concern in electronic markets.

Front-running (or prehedging) traditionally arises when a broker who is charged with executing a large order on behalf of a client trades on his own account ahead of the client trades, selling back or buying back to the client at worse prices (from the client's perspective). Nowadays, this refers to the practice whereby a trader learns about a large order through the electronic marketplace and trades ahead of that order, thereby profiting from the momentum induced by the clients trades. One way of learning about the presence of large order is through phishing (or pinging), which refers to quoting or placing orders, usually in small size, to uncover the hidden orders or intentions of other participants and then trading to take advantage of the information obtained. This technique is particularly effective on trading platforms where order confirmations are sent immediately, but market data updates are sent afterwards or even at regular, low-sampled intervals.

Quote stuffing involves placing a large number of orders and then immediately canceling them. The purpose of this is to make things difficult for rival firms by adding lots of noise into the system that the stuffer knows is not real but that other participants do not. There are claims that this happened during the Flash Crash and at other times. This seems to be a particularly serious issue in the US stock market system due to the large number of trading venues and the logic of the National Market System, which requires routing of orders to obtain the best price, wherever that is

(Elder 2010). On the London Stock Exchange, for example, there are message-throttling schemes to prevent overmessaging relative to some prior agreed quantity. Of course, in times of market crisis when trading volumes increase massively relative to average daily volume, these limits may be relaxed by the trading venue.

Smoking, flashing, or strobing refers to offering attractive limit orders (better than the current top of book) and then quickly amending these orders to worse prices to exploit slower participants' market orders.

Layering or spoofing refers to creating a series of visible limit orders on one side of the market (the fake side) entered in a sequence to create the impression of increased demand or supply but far enough away from the touch to have small execution risk (or, if close to the touch, small enough in size for execution not to matter). This is followed or preceded by a limit order away from the touch on the other side (or a market order at the right time). For the spoofing to work, other participants have to improve their quotes in response to the fake-side activity, thereby creating a better price on that side, which is then hit by the spoofer's other-side orders. This is then followed by cancellation of the fake-side orders and, perhaps, reversal of the strategy to complete a round trip. The Dodd-Frank Act prohibits disruptive trading practices such as spoofing.

The academic literature on modern market manipulations is relatively thin. The classic theoretical treatments of Allen & Gorton (1992) and Kyle & Viswanathan (2008) discuss the economics of manipulation. There are a number of empirical studies documenting manipulation of closing prices (banging the close) (see Putniņš 2012), but relatively few documenting, say, spoofing. One exception is the work of Lee et al. (2013), who investigate the Korean exchange ELOB, which until the end of 2001 disclosed the best prices and the total quantity without regard to where that quantity was priced. The authors document spoofing strategies on this market: big orders placed away from the touch that conveyed the impression of increased activity on one side of the market followed by market orders on the other side that captured the subsequent price movement.

Several market manipulation cases in the United States and the United Kingdom have led to legal action. Navinder Sarao (the Hound of Hounslow) was convicted in February 2017 of spoofing in the Emini futures market, including during the Flash Crash. He was trading from his parents' house in West London through a US broker's account. Similarly, two recent cases in the United Kingdom involved traders who spoofed stocks on the London Stock Exchange using a combination of algorithmic execution and manual intervention and were relatively small and technologically mediocre players. The relative lack of HFT-related cases is consistent with the interpretation that HFT is not giving rise to more abuse, or alternatively, that such abuse is much harder to detect. It is certainly the case that the few penalties related to HFT pale into insignificance when compared to the fines imposed to date on the firms implicated in the low-frequency London Inter-bank Offered Rate fixing scandal uncovered in 2012 and the FX fixing scandal uncovered in 2014. There is some evidence (Aitken et al. 2012) that closing price manipulation has reduced due to the presence of HFT.

8. BIG DATA AND FINANCIAL MARKETS

The amount of data created and consumed by humanity has increased exponentially in recent times and will continue to do so in the near future. Recent developments in machine learning technology promise to be able to analyze these data rapidly and accurately. The success of the computer program AlphaGo in beating the world Go champion in 2016 demonstrated the value of machine learning techniques. Go was considered the most complex and subtle of human games, more difficult for a computer to win at than chess, which was already conquered 10 years ago. In this case, the data are quite complex in type although rather small in volume, but the strategy space over which the software has to search is huge, as there are approximately $361! = 10^{700}$ different games on a 19×19 board. This new data mining methodology has achieved many other successes in medicine, marketing, and security services. Many of the basic techniques used in machine learning are easy to acquire and deploy and are publicly available in open source depositaries such as GitHub.

The objective of HFT is to make many round-trip trades in as short a time as possible with small expected profit per trade, which involves making very short-term predictions. The approach of high-frequency traders is a bit like taking AlphaGo and requiring it to make a move every nanosecond. In that case, it would make very simple moves that would not have any educational value ex post. It would still win if the human opponent was required to also move in nanosecond frequency or forfeit their move, but it would be an ugly win. Some algorithmic traders, however, are trying to buy or sell a large quantity of stock and seek to do this in a way that gets the best price within some fairly long time frame. This is an objective that may be better achieved using more complicated techniques based on big data. The number of financial technology companies that provide such services is increasing. Could we ever expect big data techniques to be able to predict financial markets better than was possible in the past?

Economic models usually employ a dichotomy between informed and uninformed individuals and draw various conclusions from this. Grossman & Stiglitz (1980) develop a model of stock prices in the presence of costly information acquisition. They establish

$$1 - \rho_\theta^2 = \frac{\exp(\gamma c) - 1}{\sigma_\theta^2 / \sigma_\varepsilon^2}, \qquad 2.$$

where the signal noise ratio is $\sigma_\theta^2/\sigma_\varepsilon^2$, the cost of signal acquisition is c, γ is the risk aversion, and ρ_θ^2 is the squared correlation between the signal and the equilibrium price, which measures the informativeness of the price system. This allows some comparative statics. One could argue that the cost of acquiring basic historical data has decreased and that, generally, the quality and quantity of data have improved, although given the vast quantity of data being produced, finding the relevant information from the noisy chatter is challenging (and the cost of the most relevant and timely data may even have even increased). Suppose that $\sigma_\theta^2/\sigma_\varepsilon^2$ increases and c decreases in Equation 2. In this case, ρ_θ^2 should rise, and the price system should be more informative. However, one might imagine a more complex scenario in which both the signal and the noise increase and in which it is the relative cost of information (cost per unit of σ_θ^2) that matters, in which case the predictions are ambiguous. One thing should be clear: The Grossman & Stiglitz model sets limits on the amount of predictability that can be achieved in financial markets for any given cost-to-signal-noise ratio. In practice, there are also many issues raised about the sensitivity of the market system to herding induced by automated systems.

9. CONCLUSIONS

HFT can improve the quality of markets, fostering greater liquidity, narrowing spreads, and increasing efficiency. Yet these benefits may come with associated costs: The rates at which current systems can interact autonomously with each other raise the risk that rare but extreme adverse events can be initiated and then proceed at speeds very much faster than humans can comfortably cope with, generating volumes of data that can require weeks of computer-assisted analysis by teams of skilled analysts before they are understood. Although they may happen only very rarely, there is a clear danger that very serious situations can develop at extreme speed. In this section, we discuss a number of proposals to alter market design to mitigate some of the negative outcomes associated with HFT.

Some have argued that the order priority rules that determine the sequence in which submitted orders are executed on equity markets are at fault and prioritize speed. The policy issue is whether time-price priority unduly rewards high-frequency traders and leads to overinvestment in an unproductive technology arms race (Baron et al. 2014, Haldane 2011). The greatest benefit of a time-price priority rule is that it treats every order equally. Using other priorities, such as a pro rata rule, where every order at a price gets a partial execution, gives greater benefits to large traders over small traders. In addition, time-price priority provides a stronger incentive to improve the quote than does a pro rata rule, enhancing liquidity dynamics. Limit order providers face risks in that traders with better information can profit at their expense. Time-price priority encourages risk taking by giving priority in execution to limit order providers willing to improve their quotes. The IEX exchange was established to offer equity traders an alternative priority scheme, as discussed by Lewis (2015). This exchange also claims to offer protection from crumbling quotes, i.e., to protect orders from trading during unstable, fast-moving, and potentially adverse conditions.[8] The IEX has gained some market share but is still far behind the NYSE, Nasdaq, and BATS.

Budish et al. (2015) have proposed replacing the continuous ELOB with periodic auctions, which can be designed to minimize the advantage of speed and to mitigate other negative outcomes of the continuous trading model, such as manipulative strategies. The main benefit of periodic call auctions would be a reduction of the speed of trading and the elimination of the arms race for speed discussed above. The speed of trading could be controlled through the timing and frequency parameters, which could be tuned to individual and market conditions. Two issues with this proposal are fostering competition and allowing dynamic hedging, which require synchronization between securities on the same and competing venues. Many markets have auctions at the open and the close and are now introducing midday auctions, in addition to the continuous trading segment. The auction mechanism was the primary mechanism for stock trading in many markets before the ELOB arrived, and there are a number of studies from this period that document the benefits of adding the ELOB (e.g., Amihud et al. 1997).

The world's financial markets are engines of economic growth, enabling corporations to raise funds and offering investors the opportunity to achieve their preferred balance of expected risks and rewards. It is important that they remain fair and orderly. Deciding how best to ensure this, in light of the huge growth in both the uptake and complexity of HFT that has occurred in the past decade and that can be expected to continue in the next, requires careful thought and discussion.

DISCLOSURE STATEMENT

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

ACKNOWLEDGMENTS

The authors would like to thank Maureen O'Hara and Jean Pierre Zigrand for discussions and University of Cambridge and the Institute for New Economic Thinking for funding.

[8] The IEX Signal uses a proprietary model to indicate whether a given quote is unstable, meaning that the national best bid (NBB) is about to decline or the national best offer (NBO) is about to increase. When the Signal indicates that the NBB (NBO) is about to decline (increase), the Signal is on. During this time, D-Peg and Primary Peg buy (sell) orders on IEX, do not exercise price discretion, and continue resting less aggressively, thus protecting these orders from trading in unstable, potentially adverse conditions.

LITERATURE CITED

Aitken M, Harris FHd, McInish T, Aspris A, Foley S. 2012. *High frequency trading—assessing the impact on market efficiency and integrity*. Rep, Foresight, Gov. Off. Sci., London

Aït-Sahalia Y, Saglam M. 2013. *High frequency traders: taking advantage of speed*. NBER Work. Pap. 19531

Aït-Sahalia Y, Yu J. 2009. High frequency market microstructure noise estimates and liquidity measures. *Ann. Appl. Stat.* 3(1):422–57

Allen F, Gorton G. 1992. Stock price manipulation, market microstructure and asymmetric information. *Eur. Econ. Rev.* 36:624–30

Amihud Y, Mendelsohn H, Lauterbach B. 1997. Market microstructures and securities values: evidence from the Tel Aviv Stock Exchange. *J. Financ. Econ.* 45(3):365–90

Andersen TG, Bondarenko O. 2014. VPIN and the Flash Crash. *J. Financ. Mark.* 17:1–46

Angel JJ, Harris LE, Chester SS. 2015. Equity trading in the 21st century: an update. *Q. J. Finance* 5(1):1–39

Aquilina M, Diaz-Rainey I, Ibikunle G, Sun Y. 2017. *Aggregate market quality implications of dark trading*. Occas. Pap. 29, Financ. Conduct Auth., London

Arak M, Cook RE. 1997. Do daily price limits act as magnets: the case of treasury bond futures. *J. Financ. Serv. Res.* 12:5–20

Bank Int. Settl. 2017. *The sterling "flash event" on 7 October 2016*. Rep., Bank Int. Settl., Basel, Switz. **https://www.bis.org/publ/mktc09.htm**

Baron M, Brogaard J, Kirilenko A. 2014. *Risk and return in high frequency trading*. Work. Pap., Princeton Univ., Princeton, NJ

Benos E, Sagade S. 2012. *High-frequency trading behaviour and its impact on market quality: evidence from the UK equity market*. Work. Pap. 469, Bank Engl., London

Bertrand M, Duflo E, Mullainathan S. 2004. How much should we trust differences-in-differences estimates? *Q. J. Econ.* 119(1):249–75

Bessembinder H. 2003. Trade execution costs and market quality after decimalization. *J. Financ. Quant. Anal.* 38:747–77

Biais B, Woolley P. 2011. *High frequency trading*. Rep., Eur. Inst. Financ. Reg., Paris

Biais B, Foucault T, Moinas S. 2015. Equilibrium fast trading. *J. Financ. Econ.* 116(2):292–313

Boehmer E, Fong KYL, Wu J. 2012. *Algorithmic trading and changes in firms' equity capital*. Res. Pap., Financ. Res. Netw., Brisbane, Aust.

Brogaard J. 2010. *High frequency trading and its impact on market quality*. Work. Pap., Financ. Serv. Auth., London

Brogaard J, Garriott C. 2017. *High-frequency trading competition*. Work. Pap. 2014-19, Bank Can., Ottawa

Brogaard J, Hendershott T, Hunt S, Ysusi C. 2012. High-frequency trading and the execution costs of institutional investors. *Financ. Rev.* 49(2):345–69

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

Brugler JA, Linton OB. 2017. *The cross-sectional spillovers of single stock circuit breakers*. Work. Pap., Cambridge Univ., Cambridge, UK/Inst. New Econ. Think., New York

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(4):1547–621

Caivano V. 2015. *The impact of high-frequency trading on volatility. Evidence from the Italian market.* Work. Pap. 80, Comm. Naz. Soc. Borsa, Rome

Cartea A, Penalva J. 2012. Where is the value in high frequency trading? *Q. J. Finance* 02:1250014

Castura J, Litzenberger R, Gorelick R, Dwivedi Y. 2010. *Market efficiency and microstructure evolution in US equity markets: a high frequency perspective*. Work. Pap., Secur. Exch. Comm., Washington, DC

Chaboud A, Chiquoine B, Hjalmarsson E, Vega C. 2014. Rise of the machines: algorithmic trading in the foreign exchange market. *J. Finance* 69(5):2045–84

Chordia T, Roll R, Subrahmanyam A. 2011. Recent trends in trading activity and market quality. *J. Financ. Econ.* 101(2):243–63

Clarke P. 2017. Want a job in high frequency trading? Here are the pay and job prospects at 14 key firms. *eFinancial Careers*, Nov. 3. **https://news.efinancialcareers.com/uk-en/199934/want-job-high-frequency-trading-pay-career-prospects-15-key-firms/**

Commod. Futures Trading Comm., Secur. Exch. Comm. 2010. *Findings regarding the market events of May 6, 2010*. Rep., Commod. Futures Trading Comm./Secur. Exch. Comm., Washington, DC

Copeland T, Galai D. 1983. Information effects on the bid-ask spread. *J. Finance* 38(5):1457–69

Cumming DJ, Zhan F, Aitken MJ. 2012. *High frequency trading and end-of-day manipulation*. Work. Pap. 210992C, Soc. Sci. Res. Netw., Rochester, NY

Danielsson J, Shin HS. 2003. *Endogenous Risk*, *Modern Risk Management: A History*. London: Risk Books

Degryse H, de Jong F, van Kervel V. 2015. The impact of dark trading and visible fragmentation on market quality. *Rev. Finance* 19(4):1587–622

Easley D, Lopez de Prado M, 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

Elder J. 2010. Quote stuffing is one more reason not to invest in stocks. *Wall Street Journal*, Sept. 8, A18

Foresight. 2012. *The future of computer trading in financial markets*. Proj. Rep., Gov. Off. Sci., London

Foucault T, Kozhan R, Tham WW. 2016b. Toxic arbitrage. *Rev. Financ. Stud.* 30:1053–94

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

Friederich S, Payne R. 2015. Order-to-trade ratios and market liquidity. *J. Bank. Finance* 50:214–23

Glosten L. 1994. Is the electronic order book inevitable? *J. Finance* 49:1127–61

Glosten LR, Milgrom P. 1985. Bid, ask and transactions prices in a specialist market with heterogeneously informed traders. *J. Financ. Econ.* 14:71–100

Goyenko RY, Holden CW, Trzcinka CA. 2009. Do liquidity measures measure liquidity? *J. Financ. Econ.* 92(2):153–81

Gresse C. 2011. *Effects of the competition between multiple trading platforms on market liquidity: evidence from the MiFID experience*. Rep., CENCOR, Mex. City

Grossman S, Stiglitz J. 1980. On the impossibility of informationally efficient markets. *Am. Econ. Rev.* 70(3):393–408

Haldane AG. 2011. *The race to zero*. Presented at Int. Econ. Assoc. World Congr., 16th, July 8, Beijing

Hasbrouck J. 2018. High frequency quoting: short-term volatility in bids and offers. *J. Financ. Quant. Anal.* In press

Hasbrouck J, Saar G. 2013. Low-latency trading. *J. Financ. Mark.* 16:646–79

Hendershott T. 2012. *High frequency trading and price efficiency*. Rep. DR12, Foresight, Gov. Off. Sci., London

Hendershott T, Jones C, Menkveld A. 2011. Does algorithmic trading improve liquidity? *J. Finance* 66(1):1–33

Hendershott T, Riordan R. 2013. Algorithmic trading and the market for liquidity. *J. Financ. Quant. Anal.* 48(4):1001–24

Ho T, Stoll H. 1981. Optimal dealer pricing under transactions and return uncertainty. *J. Financ. Econ.* 9:47–73

Ho T, Stoll H. 1983. The dynamics of dealer markets under competition. *J. Finance* 38:1053–74

Hunsader E. 2010. *Analysis of the "Flash Crash," Date of Event: 20100506, complete text*. Rep., Nanex Corp., Winnetka, IL

Jarrow RA, Protter P. 2011. *A dysfunctional role of high frequency trading in electronic markets*. Res. Pap. 08-2011, Johnson School, Cornell Univ., Ithaca, NY

Jiang JG, Lo I, Valente G. 2013. *High-frequency trading around macroeconomic news announcements: evidence from the U.S. treasury market.* Work. Pap., Bank Can., Ottawa

Jones CM. 2002. *A century of stock market liquidity and trading costs*. Work. Pap., Columbia Univ., New York

Jones CM. 2013. *What do we know about high-frequency trading?* Res. Pap. 13-11, Columbia Bus. School, Columbia Univ., New York

Jovanovic B, Menkveld A. 2011. *Middlemen in limit-order markets*. Work. Pap. 1624329, Soc. Sci. Res. Netw., Rochester, NY

Kearns M, Kulesza A, Nevmyvaka Y. 2010. Empirical limitations of high frequency trading profitability. *J. Trading* 5(4):50–62

Khandani A, Lo A. 2007. What happened to the quants in August 2007? *J. Invest. Manag.* 5:5–54

Kirilenko AA, Lamacie G. 2015. *Latency and asset prices*. Work. Pap. 2546567, Soc. Sci. Res. Netw., Rochester, NY

Kirilenko AA, Samadi M, Kyle AS, Tuzun T. 2017. The flash crash: the impact of high frequency trading on an electronic market. *J. Finance* 72(3):967–98

Krugman P. 2009. Rewarding bad actors. *New York Times*, Aug. 2. **http://www.nytimes.com/2009/08/03/opinion/03krugman.html?_r=1**

Kyle AS, Obizhaeva AA. 2016a. *Large bets and stock market crashes*. Work. Pap. 2023776, Soc. Sci. Res. Netw., Rochester, NY

Kyle AS, Obizhaeva AA. 2016b. Market microstructure invariance: empirical hypotheses. *Econometrica* 84(4):1345–404

Kyle AS, Viswanathan S. 2008. How to define illegal price manipulation. *Am. Econ. Rev.* 98:274–79

Lee EJ, Eom KS, Park KS. 2013. Microstructure-based manipulation: strategic behavior and performance of spoofing traders. *J. Financ. Mark.* 16:227–52

Lewis M. 2015. *Flash Boys: A Wall Street Revolt*. New York: W.W. Norton

Linton OB, Wu J. 2016. *A coupled component GARCH model for intraday and overnight volatility*. Work. Pap. 2874631, Soc. Sci. Res. Netw., Rochester, NY

Madhavan A. 2011. *Exchange-traded funds, market structure and the Flash Crash*. Work. Pap., BlackRock, New York

Mahmoodzadeh S, Tseng MC, Gencay R. 2017. *Spot arbitrage in FX market and algorithmic trading: Speed is not of the essence*. Work. Pap. 3039407, Soc. Sci. Res. Netw., Rochester, NY

Malinova K, Park A. 2011. *Subsidizing liquidity: the impact of make/take fees on market quality*. Work. Pap., Am. Finance Assoc., Salt Lake City, UT

Markham JW. 2014. *Law Enforcement and the History of Financial Market Manipulation*. New York: M.E. Sharpe

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

Menkveld AJ, Yueshen BZ. 2017. *The Flash Crash: a cautionary tale about highly fragmented markets*. Work. Pap. 2243520, Soc. Sci. Res. Netw., Rochester, NY

Nanex. 2012. *Knightmare on Wall Street: what really happened, or how to test your new market making software and lose a pile of money, fast*. Rep., Nanex, Winnetka, IL. **http://www.nanex.net/aqck2/3522.html**

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

O'Hara M, Ye M. 2011. Is market fragmentation harming market quality? *J. Financ. Econ.* 100:459–74

Pagnotta E, Phillipon T. 2015. *Competing on speed*. Unpublished manuscript, New York Univ.

Putniņš TJ. 2012. Market manipulation: a survey. *J. Econ. Surv.* 26:952–67

Secur. Exch. Comm. 2010. *US Securities and Exchange Commission. Concept release on equity market structure*. News Release 34-61458, Feb. 1

Shkilko A, Sokolov K. 2016. *Every cloud has a silver lining: fast trading, microwave connectivity and trading costs*. Work. Pap. 2848562, Soc. Sci. Res. Netw., Rochester, NY

Storkenmaier A, Wagener M. 2011. *Do we need a European National Market System?* Work. Pap., Work. Pap. 1760778, Soc. Sci. Res. Netw., Rochester, NY

Subrahmanyam A. 1994. Circuit breakers and market volatility: a theoretical perspective. *J. Finance* 49:237–54

TABB Group. 2012. *US equities market 2012: mid-year review*. Rep., TABB Group, New York

Tobek O, Linton O, Noss J, Crowley-Reidy L, Pedace L. 2017. *The October 2016 sterling flash episode: when liquidity disappeared from one of the world's most liquid markets*. Work. Pap., Bank Engl., London

Zhang F. 2010. *High-frequency trading, stock volatility, and price discovery*. Work. Pap. 1691679, Soc. Sci. Res. Pap., Rochester, NY

Annual Review of Economics

Volume 10, 2018

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://www.annualreviews.org/errata/economics