Journal of Financial Economics 116 (2015) 257–270

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

High frequency market microstructure ☆

Maureen O'Hara *

Johnson Graduate School of Management, Sage Hall Cornell University, Ithaca, NY 14853, USA

ARTICLE INFO

Article history:
Received 26 March 2014
Received in revised form
1 August 2014
Accepted 26 August 2014
Available online 29 January 2015

JEL classification:
D02
D83
G14

Keywords:
High frequency trading
Market microstructure
Algorithmic trading

ABSTRACT

Markets are different now, transformed by technology and high frequency trading. In this paper, I investigate the implications of these changes for high frequency market microstructure (HFT). I describe the new high frequency world, with a particular focus on how HFT affects the strategies of traders and markets. I discuss some of the gaps that arise when thinking about microstructure research issues in the high frequency world. I suggest that, like everything else in the markets, research must also change to reflect the new realities of the high frequency world. I propose some topics for this new research agenda in high frequency market microstructure.

1. Introduction

Markets are different now in fundamental ways. High frequency trading (HFT) has clearly made things faster, but viewing the advent of HFT as being only about speed misses the revolution that has happened in markets. From the way traders trade, to the way markets are structured, to the way liquidity and price discovery arise – all are now different in the high frequency world. What is particularly intriguing is the role played by microstructure. One could expect that when things are fast the market structure becomes irrelevant, but the opposite is the case. At very fast speeds, microstructure takes on a starring role.

To understand this evolution of the market from human involvement to computer control, from operating in time frames of minutes to time scales of microseconds, it is important to recognize the role played by strategic behavior. High frequency trading is strategic because it maximizes against market design, other high frequency traders, and other traders. HFT strategies can be quite complex, but so, too, are the strategies that other traders elect, in part because they need to optimize in a market that contains HFT players. And the exchanges act strategically as well, opting for new pricing models and market designs to attract (and, in some cases, deter) particular volume to their trading venues. As a result, trading has changed, and the data that emerge from the trading process are consequently altered.

In this paper, I investigate the implications of these changes for high frequency market microstructure. My goal is not to explain high frequency trading per se, but rather to set out some important aspects of this high frequency transformation. For finance researchers more

☆ My thanks to Bill Schwert and Ken French for suggesting this paper to me. I am very grateful to two referees, Ayan Bhattacharya, David Easley, Frank Hatheway, Joel Hasbrouck, David Meitz, Pam Moulton, Gideon Saar, and Mao Ye for help and guidance with this paper. I particularly wish to thank Jamie Selway, Jeff Bacidore, Wenjie Xu, Cindy Yang, and Lin Jiang for all of their help with this project. Maureen O'Hara is chairman of the board of directors of Investment Technology Group (ITG), a global broker-dealer firm focusing on the needs of buy-side institutional clients.

* Tel.: +1 607 255 3645; fax: +1 607 254 4590.
E-mail address: mo19@cornell.edu

http://dx.doi.org/10.1016/j.jfineco.2015.01.003
0304-405X/© 2015 Elsevier B.V. All rights reserved.

generally, understanding how markets and trading have changed is important for informing future research. For microstructure researchers, I believe these changes call for a new research agenda, one that recognizes how the learning models used in the past are lacking and why traditionally employed empirical methods might no longer be appropriate. Equally important, microstructure research must provide more policy guidance, reflecting the problem that the new complexity of markets can confound even the best-intentioned regulators.

Some of this agenda for high frequency market microstructure research is well under-way, with a large and vibrant literature developing on high frequency trading. In this paper, I highlight some of these new directions but stop far short of surveying the high frequency trading literature (more extensive reviews are Biais and Wooley, 2011; Jones, 2012; Goldstein, Kumar, and Graves, 2014). Instead, my hope is to demonstrate how markets have changed, illustrate the new range of issues confronting researchers, and suggest some fundamental questions I believe need to be addressed in microstructure research.

This paper is organized as follows. Section 2 describes the high frequency world, with a particular focus on how HFT affects the strategies of traders and markets. I set out some basics of the present market structure and discuss the role regulatory change played in setting the stage for high frequency trading. I consider the behaviors and strategies of high frequency and non-high frequency traders, and I examine how HFT has affected the organization of trading, giving particular attention to exchange pricing models, order priority rules, and the development of new trading platforms. Section 3 discusses gaps that arise when thinking about microstructure research issues in the high frequency world and proposes some topics for this new research agenda in high frequency market microstructure. Section 4 identifies some of the complex regulatory and policy issues needing further study in high frequency markets.

2. The high frequency world

Over the last decade, the forces of technology, speed, and computer-based trading have increasingly shaped the structure and behavior of markets. While much has been made of the activities of high frequency traders, the behavior of non high frequency traders is also now radically different and so, too, are the markets in which this trading occurs. In this section, I describe this new high frequency world, with the goal of conveying at least partially the sea change that has transformed trading.

2.1. The setting

The technology that allowed for high frequency trading was developing over the 1990s, but it was regulatory policy changes intended to increase competition that ushered in the high frequency era. In the U.S., Regulation ATS (alternative trading systems; Reg ATS) in 2000 allowed for the entry of a variety of non-exchange competitors, while Regulation National Market System (Reg NMS) in 2007 set out a vision of a market composed of multiple trading venues all linked together via rules over access and trade priority. In Europe, MiFiD in 2007 had a similar effect in allowing new competition and trading venues. As a consequence, equity markets in the U.S. and Europe fragmented, with trading dispersed across a variety of exchanges and markets.

The U.S., there are 11 equity exchanges, and 50 or more alternative trading systems (these include crossing networks and dark pools operated by broker-dealer firms) executing trades for customer orders. There are also dozens of trading desks executing trades internally at firms such as Goldman Sachs, Credit Suisse, Citibank, and the like. Added to this are 13 US options exchanges trading equity derivatives, as well as several futures markets trading relevant equity-linked contracts.

Fragmentation introduces a variety of complexities into the trading environment. Without a central market, traders need to search for liquidity across many venues and the ability to do so at high speeds is valuable. Multiple venues executing trades also means that prices need not always be the same, opening the door for arbitrage across markets. Advances in technology allowed this to happen but so, too, did decisions by exchanges to allow (for a fee) some traders to trade faster by co-locating their trading systems at the exchange site. The exchanges also offered (for a fee) direct feeds of their trading information, giving high speed traders an ability to see the market with more clarity than traders receiving standard consolidated tape data.

Such clarity is useful both for trading within equity markets, where high frequency traders can know prices in the various scattered markets before they are reflected in the slower tape, and for trading in correlated markets such as futures. The ability of high frequency traders to enter and cancel orders faster than everyone else also makes it hard to discern where liquidity exists across the fragmented markets. This uncertainty, in turn, creates even more opportunity for high frequency traders to exploit profitable trading opportunities both within and across markets.

The current market structure is thus highly competitive, highly fragmented, and very fast. It is also dominated by the trading of high frequency traders, who by some estimates make up half or more of all of trading volume. Understanding what high frequency traders do is crucial for comprehending why markets today are so very different from times past.

2.2. High frequency traders

High frequency trading is a misnomer, a seemingly precise term used to describe a large and diverse set of activities and behaviors. Certainly, all HFT activities have some things in common. HFT is done by computers, it relies on extremely fast speeds, and it is strategy-based. But within HFT, large differences can exist even in these common traits.

The HFT world breaks down into gradations ranging from low latency (very fast connections and trading speeds) to ultra-low latency (trading dependent on being at the physical limits of sending orders through time and space). Latency is the time it takes to send data (orders, messages, etc.) to a required end point (and potentially back again). The time scales involved are astonishingly small. Order latencies are measured in milliseconds (one-thousandth of a second), microseconds (one millionth of a second), and in some settings even nano-seconds (one-billionth of a second). By way of context, it takes the human eye 400–500 milliseconds to respond to visual stimuli.

While the low latency group relies on co-location of servers within exchanges and dedicated access to trading information, the ultra-group augments this with enhancements such as the Hibernian Express (an undersea cable coming on-line in mid-2014 that would reduce roundtrip latency from London to New York to 59.6 milliseconds from 64.8 milliseconds), Perseus Telecom's new microwave network between London and Frankfurt (reducing latency to below 4.6 milliseconds from the 8.35 milliseconds using a fiber-optic network), and new micro-chips capable of sending trades in 740 billionths of a second. It can be hard to fathom why such tiny differences in latencies can matter (or be worth the expense of building undersea cables, microwave towers, and the like) until one considers that Securities and Exchange Commission (SEC) data show that 23% of all canceled orders, and 38% of all canceled quotes, occur within 50 milliseconds or less of placement.[1]

Most high frequency traders want to be at the front of the queue when an attractive order arrives, and to do so requires paying careful attention to the rules and structure of the market. At a minimum, this requires maximizing one's trading strategy against a particular market's matching engine. The matching engine receives the orders sent to the exchange and determines their priority of execution. The matching engine also processes messages regarding the arrival, execution, and cancellation of orders. These messages are sent to and from the exchange as part of complex dynamic trading strategies, and it is now common for upward of 98% of all orders to be canceled instead of being executed as trades.

Most exchanges allow high frequency traders to choose from multiple latencies when connecting to the exchange's matching system. On the Tokyo Stock Exchange (TSE), latency for Arrownet, the standard service, is on the order of several milliseconds. The TSE's priority service allows users to put devices at data center entry points for the TSE network, lowering connection speeds to 260 microseconds HFTs can lower latency even further, to 15.7 μs, microseconds by co-locating trading devices at the TSE's primary site. The Exchange charges higher fees for lower latency access paths to the matching engine.[2]

Exchanges use different priority rules to sequence orders. The most common rule in equity markets is price-time priority. Orders with the best price trade first, and among those with the same price, the first order to arrive has priority. Other priority rules do exist, however. Price-size-time priority favors those willing to trade larger sizes. Many futures markets and some equity crossing networks, including ITG Posit and Morgan Stanley's MS Pool, use pro rata matching, in which all orders at a given price trade proportionately against an incoming order. It should not be surprising that priority rules affect the strategies of HFTs and other traders.

HFTs pursue a wide variety of strategies, ranging from market making activities to more pernicious trading gambits. There is general, but not universal, agreement that HFT market making enhances market quality by reducing spreads and enhancing informational efficiency (see Jones, 2012; Brogaard, Hendershott, and Riordan, 2013; Carrion, 2013). Menkveld (2013) provides direct evidence of how the entrance of HFTs reduced spreads for Dutch stocks trading on Chi-X Europe.

HFT market making differs from traditional market making in that it is often implemented across and within markets, making it akin to statistical arbitrage. Conceptually, HFT market making uses historical correlation patterns in price ticks to move liquidity between securities or markets. To understand how this is done, consider market making within a market. If statistically an upward price tick in stock A is generally followed by a similar upward price tick in stock B, then a high frequency market maker would want to sell stock A and buy stock B (essentially striving to buy low/ sell high). This involves submitting an order at the ask in stock A and at the bid in stock B. The process becomes far more complex when it goes across markets. Consider, for example, market making in exchange traded funds (ETFs). Berman (2014) highlights the case of the GLD SPDR, an equity ETF linked to gold. The HFT market maker would be quoting both in GLD and in Gold futures to take advantage of price deviations. But there are 13 other exchange traded products tied to gold, and these prices, too can diverge from gold and from each other. So this requires placing bids and asks across all of these 91 potential pairs, in the gold future and possibly in the cash market.

These orders generally (but not always) are limit orders, meaning that the high frequency trader is supplying liquidity just as in more traditional market making. But unlike its traditional counterpart, the high frequency market maker is on only one side of the book in each stock, and there is no commitment to provide liquidity continuously (see Virtu Financial, Inc., 2014 for empirical evidence on high frequency market making).[3] This has led to concerns that HFT market making can induce market instability in the guise of periodic illiquidity (see Kirilenko, Kyle, Samadi, and Tuzun, 2011; Easley, Lopez de Prado, and O'Hara, 2012; Madhavan, 2013).

Other HFT strategies employ more complex opportunistic algorithms. Some strategies are fairly straightforward, such as exploiting the deterministic patterns of simple algorithms such as TWAP (time-weighted average

[1] See Data Highlights 2013-05: The Speed of Equity Markets (October 9, 2013) and Data Highlight 24-02: Equity Market Speed Relative to Order Placement (March 19, 2014) at http://www.sec.gov

[2] See TSE Connectivity Services, at http://www.tse.or.jp/system/connectivity/index.html

[3] Virtu Financial, Inc., a leading high frequency market making firm, recently filed an S1 in advance of its proposed initial public offering. The data there show that for the period January 1, 2009 – December 31, 2013 (a total of 1238 trading days) the firm had only one day in which it lost money (see 100 of Virtu Financial, Inc., 2014).

pricing). Other strategies are more devious such as momentum ignition strategies designed to elicit predictable price patterns from orders submitted by momentum traders. Yet other strategies exploit latency differences between venues. Latency arbitrage can arise for a variety of reasons, some technological (the proprietary data feed can be faster than the consolidated feed) and some operational (co-location can allow some traders faster access, as can the ability to use new, custom order types on exchanges).

Some HFT strategies cross the line into unethical behavior.[4] O'Hara (2011) shows how a predatory algorithm can manipulate prices by tricking an agency algorithm (i.e. a broker algorithm implementing customer trades) into bidding against itself. Such a strategy can yield immediate profits (the high frequency trader sells at the now higher price) or more circuitous returns (the high frequency trader trades in a crossing network at the now higher mid-quote price). In either case, this predatory strategy is a form of spoofing and is forbidden under the 2010 Dodd-Frank Wall Street Reform and Consumer Protection Act. Egginton, Van Ness, and Van Ness (2011) and Ye, Yao, and Jiading (2013) examine quote stuffing, which involves sending and instantly canceling massive numbers of orders with the designed purpose of slowing down trading for rival HFT firms. Such manipulative behavior is also illegal.

Market participants are becoming increasingly aware that HFT endeavors are best separated into good activities and predatory activities. With this awareness have come efforts to attract good HFTs to (and deter bad HFTs from) markets through market design changes. Similarly, non-high frequency traders have become far more sophisticated, seeking out both trading strategies and trading venues that protect their interests. This, in turn, has greatly reduced the profits of HFTs. Getco, one of the largest HFT firms, saw its profits decline from $440 million in 2008 to $50 million in 2012, while for the industry overall, estimated profits have declined from roughly $5 Billion in 2009 to just over $1 Billion in 2013.[5] These declines highlight an important feature of market ecology – the rest of the market does not stand still in the face of change.

2.3. *Non-high frequency traders (i.e. everybody else (EE))*

In Section 2.2, I note that HFT is done by computers, relies on extremely fast speeds, and is strategy-based. What might not be fully appreciated is that this also describes non-HFT trading. Non-HFT trading is trading by everybody else and it includes both institutions and retail traders. All trading is now fast, with technological improvements originally attaching to HFTs permeating throughout the market place. Latencies at broker/dealer firms, the main pathway for everyone else's trading, are now below one millisecond ranging down to 500 microsecond for a market order sent directly to the exchange. Such speeds were unheard of for even HFTs a few years ago. The bottom line is that trading is now very fast for everyone in the market.

Trading is also done by computers, with algorithmic trading the mechanism for virtually all trading in markets. Algorithms are simply computer-based strategies for trading, and they are used to minimize transactions costs. Even without the complications introduced by HFT, trading is a challenging task in fragmented equity markets. Finding, and accessing, liquidity generally requires routing orders to multiple locations, all the while being cognizant of different trading fees, rebates, and access charges in each venue. Moreover, because trading patterns differ across the day, so, too, do spreads and the price impact of trades, requiring traders to optimize trading temporally as well. Add in opportunistic HFTs who spot (and take advantage of) deterministic trading patterns of unsophisticated traders, it is little wonder that EE trading now relies on increasingly sophisticated trading algorithms.

Table 1 gives a sample of the main algorithms currently offered by a large global broker-dealer firm to its buy-side institutional clients.[6] These algorithms fall into general categories of dark aggregation (relying on crossing networks and other non-displayed order strategies), scheduled (chopping orders up deterministically), volume participation (varying trading amounts to be a particular percentage of the market), active (strategic trading designed to minimize implementation shortfall), and smart (liquidity seeking dynamic trading strategies across markets and at the open and close). An interesting feature of most trade algorithms is that they are rarely pure strategies – most algorithms, for example, both supply and demand liquidity, and they typically transact across a variety of market venues.

This strategic trading by everybody else impacts markets in a variety of ways. Dark trading has become more important, trade sizes have fallen dramatically (the average trade in US equity markets is now just over two hundred shares) and odd lot trades are upward of 20% of all trades (see O'Hara, Yao, and Ye, 2014). It is important to stress that these changes are not just driven by market fragmentation. Futures markets are not fragmented, but the average trade in Treasury Bond Futures is now below 13 contracts and in WTI Crude Oil Future it is only 1.2 contracts. These small trade sizes reflect the influence of HFTs: because "silicon traders" can spot (and exploit) human traders by their tendency to trade in round numbers, all trading is converging to ever smaller sizes and it is being hidden whenever possible.

One could expect retail traders to fare poorly in this environment, but this misses the reality that retail trading has also changed. A large fraction of US retail trades are either directly internalized or delivered via purchased

[4] See Biais and Wooley (2011) for discussion of such strategies, and Jarrow and Protter (2012) for a model of high frequency manipulative strategies.

[5] Profitability figures are from Business Week, April 1, 2014. Figures on Getco profitability come from its S4 statement filed in connection with its proposed 2012 initial public offering.

[6] Buy-side clients are large asset management firms such as Vanguard, Fidelity, or T. Rowe Price as well as pension funds such as CALPERS (California Public Employees Retirement System) or the Teacher Retirement System of Texas.

Table 1
Typical trading algorithms for equity traders.
This table gives a sample of the algorithms used by customers of a large broker-dealer firm. Source of the information: ITG Alogrithms at http://www.itg.com/marketing/ITG_Algo_ExecutionStrategies_Guide_20130701.pdf.

General type	Description	Uses
Opportunistic	Posit marketplace	Has access to dark liquidity in Investment Technology Group (ITG), POSIT, and other dark venues
	Raider	Operates strategically across both dark and lit markets to capture liquidity; does not display in lit markets
	Float	Seeks to earn the spread by actively posting on the passive side
	Pounce	Has opportunistic, liquidity-seeking, finding posted and reserve liquidity and employing pegged orders to wait for liquidity in illiquid stocks
	Flex	Has customized algorithms
Implementation shortfall	Active	Dynamically trades to reduce implementation shortfall for single stocks
	Dynamic implementation shortfall	Dynamically trades to reduce implementation shortfall for baskets of securities
Participation-based	Dynamic close	Trades into the closing auction using an optimization to improve performance versus close benchmark
	Dynamic open	Optimizes participation in the opening auction
	Flexible participation	Trades using a scaling minimum and maximum participation rate relative to a benchmark and style
	VWAP	Uses predicted volume profiles to target volume-weighted average price
	Volume participation	Works trades across markets at a specified percentage of printed volume until order is filled or market closes
Strategic	Slimit	Uses anti-gaming technology and smart routing to minimize exposure to high frequency traders

order flow agreements to broker-dealer firms. For example, Charles Schwab's order flow currently is sold to UBS, meaning that UBS pays Charles Schwab to send all of its retail orders to UBS, which in turn executes those orders by taking the other side of the trade. Because of best execution requirements, UBS must provide at least the prevailing best bid or ask for these orders. Little retail trade goes directly to exchanges, in part because broker algorithms route it to a variety of other trading destinations first. Battalio, Corwin, and Jennings (2013) argue that these routing decisions are greatly influenced by the size of the rebates offered by the trading venues.[7]

When retail orders do go to the NYSE, they often benefit from liquidity provided by DMMs (designated marker makers) and SLPs (strategic liquidity providers), many of whom are high frequency trading firms.[8] The DMMs are akin to the specialists of times past, and each stock has one DMM. SLPs are high volume trading members who add liquidity to the NYSE in return for trading fee rebates. Trading costs of retail traders have been falling generally over the past 30 years (see Angel, Harris, and Spatt, 2011), and this decline seems to have accelerated. Malinova and Park (2013) provide empirical evidence that retail trading costs in Canada have directly fallen because of the entrance of high frequency traders.

2.4. Exchanges and other markets

With trading strategic and computer driven, the microstructure of trading venues takes center stage. Exchanges face a conundrum with respect to microstructure issues. With HFT more than half of trading volume, making its microstructure attractive to high frequency traders entices needed volume (and liquidity) to the exchange. However, becoming too HFT-friendly risks alienating EE traders (the institutional and retail traders) who could then choose to trade in specialized venues elsewhere.

This dilemma over market design is only the latest chapter in the market structure evolution that began with Reg. ATS and Reg. NMS, the SEC regulations ending the one-size-fits-all model of exchange trading by allowing new competitors to enter. These new trading venues crafted microstructures to meet the particular needs of specific traders. Existing exchanges, faced with competition from all sides, responded by creating markets within markets, setting up specialized microstructures to attract particular trading clienteles.

The end result is that trading is both fragmented and extremely fluid. Orders can be routed to a trading venue with the touch of a computer key and routed away just as swiftly. Exchanges and trading venues face intense competition to get the right order flow, avoiding if possible the toxic orders that disadvantage other traders. The key to doing so involves a variety of strategic decisions with respect to market design.

One such decision is the market's pricing structure. In electronic markets, liquidity arises from limit orders in the book. Traders who submit those orders are said to make liquidity, while traders who hit existing orders via market

[7] Purchased order flow agreements predate the high frequency era, but are widely agreed to be far more important now. Battalio, Corwin, and Jennings (2013) provide evidence that some large retail brokers sell all their order flow in purchased order arrangement, while others sell only their market orders but rout their limit orders to the venues giving the highest liquidity rebates.

[8] The DMM firms are Barclays Capital Inc., Brendan E. Cryan & Co. LLC, Goldman Sachs & Co., J. Streicher &Co. LLC, KCG, and Virtu Financial Capital Markets LLC. The SLPs in NYSE securities are Barclays Capital, Inc., Citadel Securities LLC, HRT Financial LLC, Bank of America/Merrill, Octeg LLC, Tradebot Systems, Inc., Virtu Financial BD LLC, KCG, and Goldman Sachs &Co. See also O'Hara, Saar, and Zhang (2013) for analysis of DMM and SLP participation in stocks.

orders are said to take liquidity. Island ECN in 1997 introduced maker-taker pricing in which market order traders paid trading fees while limit order traders received rebates, and this is now the dominant pricing model in equity markets. Such a pricing framework is particularly attractive to high frequency traders who with their speed can submit (and cancel) limit orders before everyone else, making limit order trading less risky for them. The rebates from trading via limit orders in maker-taker markets are a substantial source of profit for high frequency traders.

But maker-taker is not the only way to structure a market. There are traditional venues in which both sides of a trade pay a trading fee, taker-maker markets in which rebates accrue to the market order providers and fees attach to the limit order submitters, and even subscription markets in which you can trade as much as you want for a given monthly fee.[9] The taker-maker pricing strategy harkens back to the notion that certain orders, for example, retail flow, is less toxic (i.e. less information related) and so is more desirable to transact against. Payment for order flow was one way to attract such flows to a market, and taker-maker pricing can be thought of as a variant of that pricing strategy.[10] Ye and Yao (2014) argue that taker-maker pricing also provides a way for non-high frequency traders to jump to the head of the limit order queue by paying the maker fee. In general, taker-maker venues are thought to be less attractive to HFTs.

Because markets feature multiple trading platforms, a trading venue can attract some clienteles to one platform and different clienteles to another (or even the same clientele pursuing different strategies on each platform). For example, the BATS Global Markets features four trading platforms BZX, BZY, EdgX and EdgA. The BZX and EdgX platforms feature maker-taker pricing, and BZY and EdgA feature taker-maker. The BZX features an interesting variant on maker-taker pricing by scaling the rebate depending upon whether the maker sets, joins, or is outside the NBBO (the national best bid or offer). This differential rebate enhances market quality by incentivizing liquidity provision at the current bid and offer.

Exchanges use different order types to appeal to high frequency traders. For example, Direct Edge introduced Hide not Slide orders, a complex order type allowing submitters to circumvent rules designed to prevent locked markets (a market is locked when the ask is equal to the bid).[11] The queue-jumping feature of these orders elicited complaints that they unfairly disadvantage other traders. An alternative view is that these orders allow exchanges to compete with the algorithmic capabilities of broker-dealers by providing traders an enhanced ability to control the execution of their orders. Thus, exchanges now face competition not only from other exchanges but also from broker–dealer firms.

Trading venues also compete via access and speed. Nasdaq's new venture with Strike Technology sends data between Nasdaq's New Jersey data center and the Chicago Mercantile Exchange's data center in Aurora, Illinois, in 4.13 milliseconds. The NYSE's proposed venture with Anova Technology using laser-millimeter wave technology could be even faster. For high frequency traders, these technological innovations are key to deciding where to trade; for exchanges and markets, these innovations are key to their competitiveness (and survival). Whether society is enhanced by such an arms race in technology is debatable (see Brogaard, Garriott, and Pomeranets, 2013; Cespa and Vives, 2013; Haldane, 2012; Pagnotti and Phillipon, 2013; Biais, Foucault, and Moinas, 2012; and Budish, Cramton, and Shim, 2013). What is not in question is how expensive these technologies are. Laughlin, Aquirre, and Grundfest (2012) estimate that a 3 millisecond decrease in communication time between Chicago and New York markets cost a staggering $500 million.

Conversely, markets designed to limit the involvement of HFTs are another dimension of strategic competition. The IEX in the US and Aequitas in Canada designed microstructures to protect EE traders from HFTs. The IEX is a dark pool featuring price-broker-time priority, so orders from an agency broker have higher priority than orders coming from a high frequency trader. Orders in this market are also slowed down by going through a coiled cable that adds a 350 microsecond time delay, a feature designed to negate any speed advantage to the high frequency traders. Aequitas is not yet in operation, but its proposed modus operandi is similar. Aequitas features a matching scheme of price, broker, market maker, and weighted size-time priority. There is no maker-taker pricing, but the microstructure gives priority to market makers to incentivize liquidity provision.

A within-market approach to accomplish a similar goal for retail traders is the NYSE's Retail Liquidity Program (RLP). In this program, retail orders are submitted to the exchange by retail member organizations (they cannot be sent by algorithms or any computer methodology) and these orders execute against liquidity provided by designated retail liquidity providers (who must be NYSE designated market makers or supplemental liquidity providers). Executed trades receive price improvement relative to the best bid or offer of at least 0.001, with RLPs quoting in increments of 0.001 to provide this liquidity.

An interesting feature of this program is that the retail liquidity providers include high frequency firms such as Citadel Securities, Octeg LLC, Tradebot Systems, Inc., and Virtu Financial BD LLC. Thus, retail orders trade directly with high frequency traders who, in this taker-maker market, pay a fee for interacting with the retail flow. It remains to be seen if this market design succeeds in luring retail trade away from other trading venues. One thing it

[9] All you can eat subscription pricing is offered by the Aquis Exchange, a new pan-European trading platform that began trading in November 2013.

[10] The options exchanges feature a split between markets that use maker/taker pricing (such as the Nasdaq and BOX) and those that use payment for order flow models (such as the Chicago Board Options Exchange and International Securities Exchange). Battalio, Shkilko, and Van Ness (2011) provide an interesting analysis of the effects of these option market pricing schemes on execution quality.

[11] In locked markets, orders must move (or slide) to a worse price that unlocks the market. The hide not slide feature allowed such orders to be hidden, and not slide. These hidden orders then reverted to regular limit order status when the market unlocked, but with the advantage of being first in the queue at that price.

has attracted is competition from competing venues. BATS set up a similar retail-focused program called RPI.

The high frequency world thus constantly evolves. New technology and greater speed lead to new strategies, which lead to new methods of trading and to new market designs. Hidden within this new paradigm are other changes such as the evolving nature of liquidity, the changing character of information and adverse selection, and transformations of the fundamental properties of market data such as buys and sells, quotes, and prices. These changes, in my view, are equally important because they challenge the ways researchers interpret market data and analyze market behavior and performance.

3. Microstructure research: what is (or should be) different?

With markets and trading radically different, myriad questions demand the attention of researchers. These issues run from the particular (how do specific trading strategies affect market performance), to the general (how has market quality fared in this new environment), to the conceptual (how should markets be designed and what activities should regulation allow?) While acknowledging the importance of this research, I believe that HFT has also altered some basic constructs underlying microstructure research. In this section, I consider these more basic issues, with a goal of setting out some fundamental issues I believe are no longer well captured by the existing models and approaches.

3.1. Information in a high frequency world

Learning is an important feature in many microstructure models.[12] In their canonical form, such models rely on a basic story: some traders have private information and they trade on it; other traders see market data and they learn from it; and market prices adjust to efficient levels that reflect the information. Microstructure enters by influencing the types of market information traders see and the ease which they can learn from it. In this process, trades play a particularly important role. Buy trades are viewed as noisy signals of good news; sell trades are noisy signals of bad news. Traders (and the market) also learn from data such as orders, trade size, volume, time between trades, etc. This linkage between the learning of traders and the efficiency of markets is one of the major contributions of modern microstructure theory.

In the high frequency world, the basic story remains the same: Traders still have to trade to profit from information, and other traders still try to learn what they know from watching market data. But some things are very different. Traders are silicon, not human. Market data are not the same. Algorithmic trading means that trades are not the basic unit of market information – the underlying orders are. Adverse selection is problematic because what even is underlying information is no longer clear. How, or even what, you are trying to learn becomes a very complex process.

Consider, for example, the issue of information. Microstructure models were always vague, portraying private information as a signal of the underlying asset's true value. But in the high frequency world, it is not clear that information-based trading necessarily relates to fundamental information. This is because the time dimension that affects high speed trading also affects market makers. Whereas the time horizon of the NYSE specialist was at one point measured in weeks (see Hasbrouck and Sofianos, 1993), now it is measured in seconds, perhaps even milliseconds. Over these intervals, information might not just be asset-related but order-related as well. Haldane (2011, p.4) makes the point that “adverse selection today has taken on a different shape. In a high speed, co-located world, being informed means seeing and acting on market prices sooner than competitors. Today, it pays to be faster than the average bear, not smarter. To be uninformed is to be slow.”

This notion of speed being synonymous with informed trading is surely not the complete story, but it speaks to the complexities of information in the high frequency age. Informed trading is multidimensional in that traders can know more about the asset or about the market (or markets) or even about their own order flow and use this information to take advantage of liquidity providers. For example, markets and data providers now sell access to public information seconds (or even milliseconds) before it is seen by other traders (see Easley, O'Hara, and Yang, forthcoming). This turns public information into private information and corresponds, albeit for a very short time, into the classic information-based trading of microstructure models. But HFTs also turn speed into information via co-location and other technologies. If this allows them to predict market movements better than other traders, then they, too, are clearly informed traders.

Even large traders who know nothing special about the asset's value can be lethal to market makers simply because they know more about their own trading plans. Trade imbalances are problematic for market makers because they are always on the other side: buying if traders are selling and selling if they are buying. Trading that is heavily skewed to buys or sells is toxic and can lead market makers to withdraw from trading as their inventory or short positions reach preset parameters (recall that the market makers are algorithms, so risk management is programmed in via limits on positions).[13] Over the short time intervals of interest to market makers, even these classically uninformed traders are informed traders in the new high frequency world.

[12] In microstructure there are also inventory models such as Ho and Stoll (1981) or Foucault, Kadan, and Kandel (2005) and search-based models (see Duffie, Garleanu, and Pedersen (2005) that generally eschew information issues.

[13] Virtu Financial, Inc. (2014) stresses that this behavior is key to risk management, noting that “in order to minimize the likelihood of unintended activities by our market making strategies, if our risk management system detects a trading strategy generating revenues outside of our preset limits it will freeze, or lockdown, that strategy and alert risk management personnel and management.” (See pp. 2 and 109 of its prospectus.)

From a research perspective, these expanded definitions of informed trading are worrisome. Now, it is not clear what drives the adjustment of prices or, more to the point, where they are going. Analyses of market efficiency suggest that markets generally remain informationally efficient, which should allay at least some concerns for asset pricing researchers. But episodic instability is now characteristic of markets, driven by the desires of the informed high frequency market makers fleeing when they suspect other more informed traders are present. Markets also are more tightly inter-connected, sewn together by market making/statistical arbitrage that operates across, not just within, markets. These characteristics suggest that liquidity factors play an increased role in asset pricing. What these liquidity factors capture, and how to even measure them, is problematic.

To understand these asset pricing issues, the new role (and definitions) of adverse selection, information, and liquidity need to be better understood at a microstructure level. The artificial divergence in microstructure models between those focusing on information and those focusing on inventory is unworkable, a victim of a world in which anything that affects inventory may be thought of as information. The fiction in microstructure models of a risk neutral single market maker (which, in turn, proxies for competitive pricing in markets) may not be accurate given that pre-set risk limits induce non-participation by silicon liquidity providers. The need for new and better microstructure models seems clear.

3.2. Market data

Finding the informed traders is a fundamental issue in microstructure models, and it speaks to the issue of learning from market data. That informed traders leave footprints in markets is well established, and it is why microstructure models attach such significance to trade data. Every trade has a buyer and a seller, but in microstructure interest has been in the active side because of its signal value to the underlying information.[14] In the past, this active side was a market order hitting the specialist quotes or crossing against a limit order on the book.

The high frequency world operates differently. Algorithms chop a parent order into child orders, and these child orders (or some portion of them) ultimately turn into actual trades. Unfortunately, neither the market nor the researcher can see these parent orders, and the child orders could have very different properties. For example, child orders are not independent, so trades are not independent either, and sequences of trades become informative (it is these patterns in trading that HFTs often try to exploit).

Dynamic trading strategies mean that these orders need not result in the simple buy and sell trades of times past. Because of maker-taker pricing, algorithms rely more on limit orders to reduce the transactions costs of trading. Algorithms also make extensive use of mid-point orders, a type of limit order that adjusts the limit order price to the moving mid-point price, and they also trade at the mid-point in crossing networks. Sophisticated traders cross the spread (i.e. buy at the ask price and sell at the bid price) only when it is absolutely necessary.

To see why this matters for interpreting market data, consider a parent order to buy five thousand shares. Whereas in the days of specialist trading this order would execute as one or possibly several market buy trades, now the algorithm turns the parent order into scores of limit orders placed in layers on the book (or across many books), with orders canceled and updated as trading progresses. Because these are limit orders, any executing order is the passive side of the trade – so this five thousand share buy order shows up in the data as many small sell trades.

Does this happen? I looked at execution data from ITG (a large broker-dealer firm) for a sample of equity executions in their standard volume-weighted average price (VWAP) algorithm in the year 2013.[15] The parent orders are for at least one round lot, are VWAP market orders (i.e., they did not specify limit prices), and are fully executed within the trading day. The sample size is 243,772 parent orders. Executed trades are classified as passive if the order is buying at or below the bid (or selling at or above the offer); as aggressive if the order is buying at or above the offer (or selling at or below the bid); and as midpoint if the order is filled at prices within the spread.[16] The VWAP algorithm operates differently depending upon factors such as the size of the order and the customer's preference over execution speed, so the data are broken out by participation rate (i.e., the order size as a percentage of volume executing over some time interval) and by order size as a percentage of the total volume executed for the day.[17]

Table 2 provides execution data on these VWAP orders. The data clearly show the transition from parent orders to child orders: the algorithm executed 13,468,847 child trades, so on average each parent order turned into 55.325 child executions. The data also show that the algorithm executes the vast majority of parent orders with passive executions. For the sample as a whole, 65.3% of trades were passive, 21.9% were midpoint trades, and 12.57% were aggressive. Thus, a parent order to buy shows up in the data at least two-thirds of the time as sell orders and, including midpoint orders, this could be as high as 87%. Less than one in eight executed trades cross the

[14] Traders informed of good news have to buy to profit on their information; traders informed of bad news have to sell to make a profit. Models such as Gloston and Milgrom (1985) and Kyle (1985) use buys and sells as inputs to the market maker's pricing problem. Buy–sell data are also an input in PIN (probability of informed trade) models (see Easley, Kiefer, O'Hara, and Paperman, 1996), and it plays a role in explaining liquidity linkages across markets (see Holden, Jacobsen, and Subrahmanyam, 2014).

[15] I thank Jeff Bacidore, Wenjie Xu, Cindy Yang, and Lin Jiang for providing the data and technical analysis.

[16] So, for example, a buy order executing at the bid is a limit buy order executing against a market sell order, whereas a buy order executing at the ask is a market buy order executing against a limit sell order.

[17] The calculations are based on dollar-value weighted averages. Share-weighted averages yield similar results.

Table 2

Volume-weighted average price (VWAP) execution data.

This table gives data from a VWAP algorithm executed for Investment Technology Group (ITG) buy-side client market orders. All parent orders are market orders and are fully filled. All parent orders have at least one hundred shares. Locked quotes are eliminated in fill aggressiveness and trade statistic calculations. Percentages given are dollar-weighted. Panel A gives data split by the participation rate (amount as a percentage of volume) of the order. Panel B gives data split by the order size relative to that day's total volume. The sample period is 2013. Source of the information is ITG.

Fill aggressiveness	Number of parent orders	Number of trades	Dollar-weighted		
Participation rate			Passive	Aggressive	Midpoint
Panel A: Participation rate bucket					
0–1%	144,121	4,112,103	77.18%	7.26%	15.57%
1–5%	69,341	4,921,553	68.27%	7.93%	23.79%
5–10%	16,161	2,268,437	61.22%	12.53%	26.25%
10–25%	10,047	1,785,376	52.18%	23.87%	23.95%
25–50%	2,627	341,559	32.9%	47.93%	19.10%
50–100%	1,475	39,819	12.42%	68.80%	18.79%
Total	243,772	13,468,847	65.53%	12.57%	21.90%
Panel B: Order size bucket					
0–1%	210,557	6,254,707	71.90%	10.62%	17.48%
1–5%	27,243	4,918,155	62.77%	11.85%	25.38%
5–10%	4,374	1,533,693	57.95%	16.49%	25.56%
10–25%	1,448	702,747	53.45%	21.19%	23.95%
25–50%	137	58,551	36.89%	43.20%	19.91%
> 50%	13	994	3.82%	89.87%	6.31%
Total	243,772	13,468,847	65.53%	12.57%	21.90%

spread and are the classic buy trades of microstructure models.

Trade size and intensity affect these numbers. For small orders, 10.62% of executions are aggressive, and this fraction of aggressive trades gradually increases with order size until reaching very large trade sizes when it accelerates. For parent orders as large as 25% (50%) of the day's dollar weighted volume, aggressive executions are still less than a quarter (half) of executed trades. For massive orders, aggressive trading can exceed passive, but these orders are a miniscule 0.0005% of the total sample. Even parent orders above 10% of the day's volume are very rare (0.06% of the sample).

Trading intensity data tell a similar story. Trades participating at low rates (below 5% of volume) cross the spread less than 8% of the time. As trade intensity picks up, aggressive executions increase, but they remain very low, in part because midpoint executions take on increased importance. Parent orders participating at rates up to 10% of the dollar-weighted volume, for example, result in child order executions of 61.22% passive, 26.25% midpoint, and 12.53% aggressive. Orders trading faster than this are more aggressive but are exceedingly rare.

That trading intentions and executed trades can be very different is important for a variety of reasons. Consider, for example, the controversy surrounding the origins of the flash crash of May 9, 2011. The SEC and Commodity Futures Trading Commission staff report identified the causal factor as a "large trader" submitting a sell order at approximately 2:00 p.m., which then caused the market to fall precipitously. However, by using the actual parent order execution data from Waddell and Reed (the large trader), Menkveld and Yueshen (2013) show that this explanation is incorrect. The order's execution involved large numbers of limit sell trades, meaning that this trader was providing liquidity to the market, not taking it.

More fundamental is whether we can actually "link" buy and sell trades with underlying information. Easley, Lopez de Prado, and O'Hara (2013) argue that if signed orders are informative, then order imbalance should be related to price changes, consistent with price adjusting to new information. If, instead, it is uninformed traders who cross the spread, then order imbalances should have little relation to price changes. A simple illustration of their argument is given in Fig. 1 which shows the relationship between the absolute value of signed trade imbalance and the spread between the high and low price over 10-minute trading intervals. The data are from the Nasdaq high frequency database and are for all trading in Apple (APPL), Intuitive Surgical, Inc. (ISRG), and NX Stage Medical, Inc. (NXTM) in October 2010.[18] The figure shows virtually no relation between order imbalance and the high-low price, consistent with the active side of the trade being related more to one's willingness to cross the spread than it is to information-based trading.

This change in information content of buys and sells should not be unexpected given the changing nature of traders' execution strategies. Retail trades, which surely correspond to the uninformed trades of microstructure models, end up crossing the spread because they are internalized and are given either the best bid or the best offer. Informed traders (who either are sophisticated traders or, if Haldane is correct, are HFTs) use dynamically changing layers of limit orders to trade, rarely if ever needing to show their hand by crossing the spread.

[18] The high frequency data set includes all trades taking place on Nasdaq for 120 selected stocks over a limited sample period. The data were divided into size terciles and then the largest stock in each tercile was selected. Thus the three stocks were selected as representative of large, medium and small stocks. The data include buy and sell indicators. I am grateful to Mao Ye for his help with this analysis.

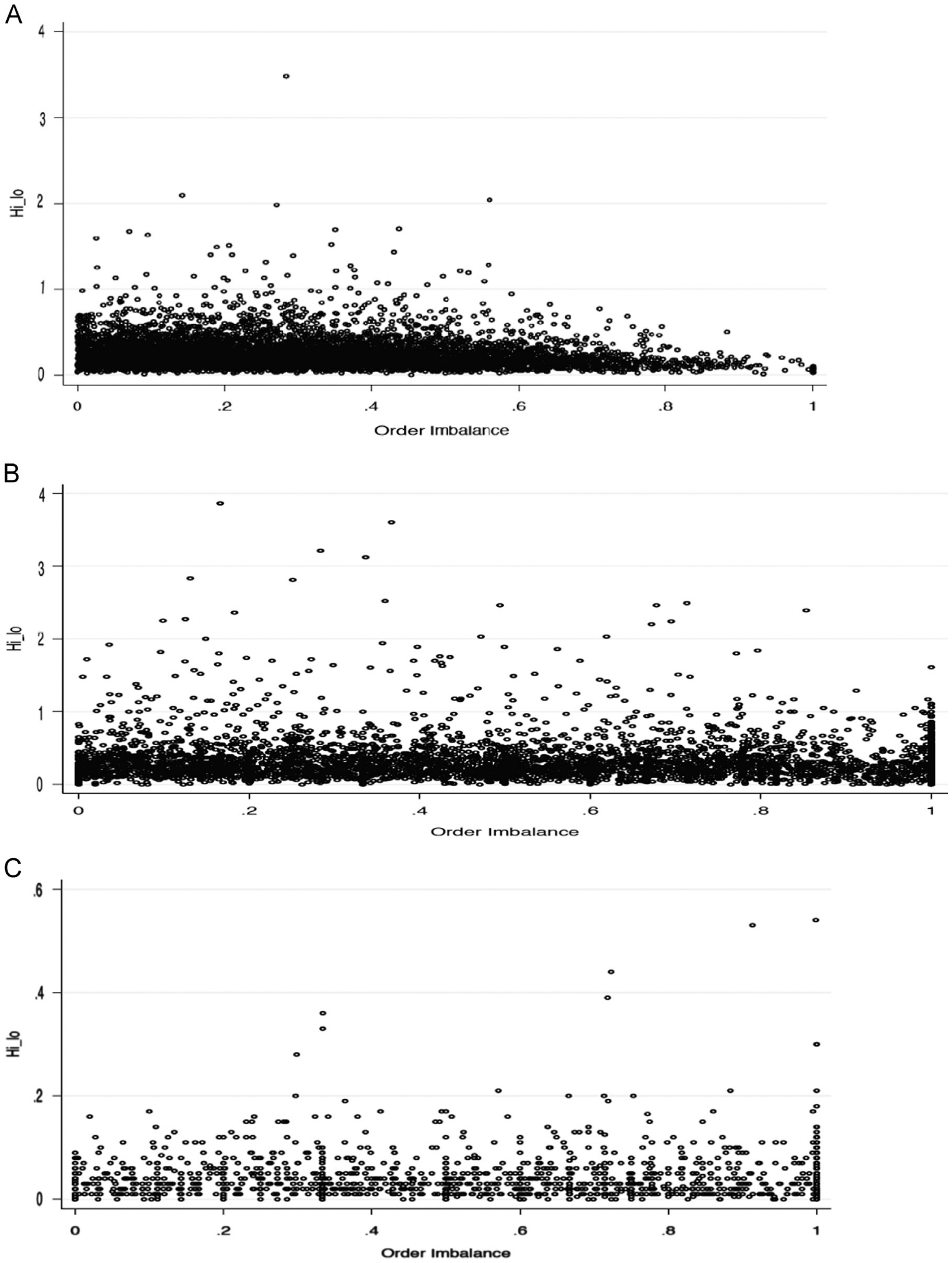

Fig. 1. Trade imbalance and price effects. This graph show the relation between the absolute value of order imbalance [(buys − sells)/(buys + sells)] and the log of the high − low price calculated over 10-minute periods. The data are from the Nasdaq high frequency data set and include each trade done on the Nasdaq exchange, excluding trades done in the opening, closing, or intraday crosses. The sample period is October 2010. Buys and sells are identified by aggressor flags in the data. Each dot presents the relation between order imbalance and high-low price in each 10-minute interval. Panel A shows trading in Apple (Large stock), Panel B shows trading in Intuitive Surgical Inc. (ISRG) (medium stock), and Panel C shows trading in NX Stage Medical, Inc. (NXTM) (Small stock).

Bloomfield, O'Hara, and Saar (forthcoming) use experimental markets to show how informed traders make more extensive use of hidden orders in exchange settings, consistent with this more nuanced world of trading.

If one cannot learn from buys and sells, what should be looked at to infer underlying information? The high frequency world gives clues in that HFT algorithms draw inferences from trade sequences and time patterns, from cancellations and additions to the book, and from volumes, to name just the obvious suspects. Exactly what these variables convey is not entirely clear, and more research is needed to ascertain what can be learned from these market data. Hasbrouck and Saar (2013) is a good example of this new research in that it highlights the role played by runs and sequences of high frequency trades in affecting market behavior and quality.

Even more important is to recognize that these data must be looked at across markets, and not just within individual markets. High frequency algorithms operate across markets, and if order books are linked, then so, too, must be order flows and price behavior. Theoretically modeling such interrelations is daunting, so empirical analyses focusing on the predictive power of market variables both within and across markets can be a good place to start. Certainly, understanding the changing nature of market data is an important direction for future research.

3.3. Analyzing data

With trading electronic, there is a wealth of trading data, and new data sets are becoming ever more available. But this treasure trove comes at a cost, both figuratively and literally. Data sets are expensive to purchase, store, and manipulate. Moreover, the massive quantities of data drawn from a variety of markets and venues pose challenges for even basic analyses of microstructure data. In the high frequency era, new tools are needed in the microstructure tool box.

Consider, for example, issues connected with the consolidated tape. In the US, all equity trades must be reported to the consolidated tape on a real time basis. Yet, the seeming precision of this requirement is illusory. Odd lots, for example, were not reported, an omission largely explained by historical conventions. O'Hara, Yao, and Ye (2014) demonstrate that odd lots play an expanded role in the high frequency world, with some stocks having 50% or more of trades execute in odd lots (the average across all stocks is around 20%). These authors also show that odd lots have high information content, consistent with informed traders using odd lots to hide trades from the market. The SEC recently required odd lot reporting to the tape, but a variety of other data, such as Rule 602 (trade execution quality) statistics, still do not include odd lots, and historical data remain incomplete. Such missing data are a natural concern to researchers.

The consolidated tape has another problem: the data could be out of order. There are 16 lit equity markets in the United States and 50 or more other trading venues. Each reports trades to the tape, but at differing latencies. The time stamps on the tape reflect when the trade report is received, not when it occurred. Equally important is that the true state of the market might not be visible to researchers using the standard monthly TAQ (Trades and Quotes) database (MTAQ) which time-stamps data only to the nearest second. Holden and Jacobsen (2013, p. 1748) provide disturbing evidence that "For MTAQ, when compared to [daily TAQ] DTAQ (which time stamps to the millisecond), we find that (1) the percent effective spread is 54% larger, (2) the percent quoted spread goes negative 37 times more often, (3) the percent quoted spread is 47% smaller, (4) the effective spread is greater than the quoted spread 15% more often (5) trades happen outside the NBBO eight times more often (6) the percent realized spread is 12% larger, and (7) the percent price impact is 109% larger."

The high frequency world also challenges empirical analyses using quote data. Quotes in microstructure traditionally represented the current price for a stock. In particular, quotes reflect the expected value of the asset given that someone wants to buy (the ask) or sell (the bid), and the midpoint of the quotes is often viewed as the current unconditional expected price. Quotes derive from actual orders on the book, but in high frequency markets the vast majority of these orders are canceled. Cancelations, revisions, and resubmission of orders all contribute to flickering quotes, creating uncertainty as to the actual level of current prices.

Hasbrouck (2013) demonstrates that this quote volatility has a number of undesirable effects such as a decrease in quote informational content, an increase in execution risk for traders, and a reduction in the reliability of the midpoint as a reference price for crossing networks. He argues that analyses of quote volatility must recognize the role of traders' time horizon, an issue I raised earlier in the context of the market maker's horizon. He proposes a new methodology employing sliding time scales, some as short as 50 milliseconds, to decompose bid and ask volatility. He demonstrates that trading induced volatility at these very short time horizons is many times larger than the volatility related to fundamental private or public information.

Time scale decomposition to facilitate analyses of high frequency market data reflects a basic reality of the high frequency world: time is not a meaningful concept in a computer-driven low latency world. Easley, Lopez de Prado, and O'Hara (2011, 2012) make a similar argument for using a volume clock in their analysis of toxicity risk in high frequency markets. Using a volume clock reduces the bias in empirical analysis arising from irregularly spaced data, a feature characteristic of high frequency markets.

These market dynamics also require care in the application of existing empirical techniques. Some favorites from the microstructure tool kit now simply do not work. Realized spreads (the difference between the trade price and the midpoint of the spread five minutes later), for example, are now often negative. Such spreads have been viewed as the returns to market making in the stock but this interpretation now seems doubtful. What causes this aberrant behavior is unclear, but it could simply reflect that in high frequency settings, five minutes is a "lifetime", and so is not a meaningful time frame in which to evaluate trading. Perhaps five seconds or fifteen seconds is a better horizon, or perhaps the realized spread is just not a useful concept anymore.

Similarly, constructs such as permanent and transitory price effects are suspect, victims of the problem of what time frame actually constitutes transitory (milliseconds? seconds?) or permanent (10 minutes? hourly? daily?).[19] Estimations using numbers of trades (such as PIN estimation) are problematic, reflecting the difficulty of estimating maximum likelihood functions when the variables must be raised to powers in the tens of thousands. Trade classification algorithms such as the Lee-Ready algorithm are undermined by a range of problems such as quote volatility, order sequence problems, and timing issues between quotes and trades.

It is tempting to believe that these issues can all be solved by better data sets, that using perfect data can fix any problem. In my view this thinking is wrong (or at best naïve), and it reflects a basic misunderstanding of the high frequency world. Data sets cannot keep up because HFT keeps evolving. Replacing monthly TAQ (MTAQ) with daily TAQ (DTAQ) as suggested by Holden, Jacobsen, and Subrahmanyam (2014) help researchers, but DTAQ with its millisecond time stamps is already challenged by trading taking place at microsecond frequencies. Knowing what is identified as a buy or sale is useless if what you want to know are the trading intentions underlying the order. Having a consolidated tape is helpful for following the market, but quote volatility (and differential latency in accessing the market) means that it tells you little about the price at which you can trade.[20]

New tools need to be developed for empirical analysis. I along with my co-authors have been working on new empirical measures of toxicity called volume weighted probability of informed trading (VPIN), as well as new empirical approaches to classify trading activity. But fundamental issues need to be addressed such as do all of these data need to be analyzed or can some sort of optimal sampling approach be used instead, or can simple nearly sufficient statistics be found? The answer to these questions depends upon what you want to know, and in the high frequency world there is no shortage of things that are not yet understood.

4. Research and the regulatory agenda

What I have not yet discussed are the many policy and regulatory issues needing research at the microstructure level. Regulation trails practice and arguably, regulators were too slow to recognize how fundamental the changes wrought by HFT were. It took the SEC almost six months to decipher the equity market piece of the flash crash, a delay in large part due to the SEC's not having the equity market data. In Europe, HFT brought to the surface long-standing problems related to market linkage across different national boundaries, difficulties greatly exacerbated by Europe's lack of any consolidated tape.

The regulators are catching up. The SEC's MIDAS (Market Information Data Analytics System) for trade surveillance uses software from Tradeworx (a leading HFT firm), giving the SEC the same tools used by the firms it is trying to regulate. Combined with the new Consolidated Audit Trail, SEC research is providing a wealth of insight into the current state of US markets (see Securities Exchange Commission, 2013, 2014).

What regulatory changes are needed for the HFT world? While regulations such as the SEC's naked access rule seem well thought out and appropriate for high frequency markets, not all regulatory efforts seem as perceptive. The transaction tax contemplated in Europe or proposals to restrict order cancellations or algorithmic usage seem ill-conceived and out-of-touch with new market realities (see Linton, O'Hara, and Zigrand, 2013). On these issues academics can play an important role by doing basic policy research. Let me conclude by highlighting two policy issues that, in my view, are particularly important. These are market linkages and the issue of fairness.

A core feature of HFT is that it spans fragmented markets. How those individual markets are tied together is crucial for determining how well the overall market functions. In the U.S., the trade-through rule requires a market not quoting the best price (i.e. either the lowest ask or the highest bid) to send any order it receives to the market that is doing so, or to match that better price itself. Just as retail stores routinely advertise “we will match any better price,” so, too, can markets compete simply by matching, instead of setting, the best price.

In principle, trade-through ensures that an order gets the best price, and it allows competing venues to coexist. But it has a variety of other effects. Internalization, for example, is made possible because large banks can execute orders on their trading desks or in their dark pools by matching the current national best bid or offer. Internalization, in turn, begets payment for order flow to retail brokers, resulting in fewer orders going to exchanges. Matching, instead of making, the best price also undermines incentives to place limit orders. Trade-through makes order routing predictable, allowing high frequency traders to step in front of orders heading to the market with the best price.

Is there a better way to link markets? It is not clear. Canada recently adopted a trade-at rule in which the market posting the best price must get the order. But this has resulted in reduced competition, concentrated trading at the TSX (the Toronto Stock Exchange), and negative effects on some measures of trading costs. It has also resulted in diverting order flow in Canadian stocks to US markets.

In my view, trade-through is not optimal for the high frequency world, but what to replace it with requires more

[19] Hendershott, Jones, and Menkveld (2013) propose a new methodology for measuring temporary price effects of an order that takes into account the complications introduced by the chopping of parent orders into myriad child orders. Their analysis illustrates the complexity of measuring trading costs in high frequency settings.

[20] It might not even be possible to have perfect data in fragmented markets because latency differences between venues mean that some traders' information sets include events before other traders can see them. This is clearly a problem with using the current time stamps on the consolidated tape (which can cause trades to appear on the tape out of sequence). This would still be a problem if reporting rules change to using the time stamps when the trades occurred as all traders will not have synchronous information on all markets. I thank the referee for raising this point.

research. Should matching be allowed only if the price is significantly improved? Is displayed price even the right metric to consider or might not priority attach to other attributes such as depth? With 30% or more of all orders on exchanges hidden, is the best price even knowable? Is the national best bid or offer still a workable concept? The regulatory issues surrounding linkages are complex and important.

A second issue is fairness. There is increasing concern that while markets are faster, they are not fairer. Fairness is not an issue typically considered in microstructure, when the focus has been on market properties such as liquidity and price efficiency. But the greater complexity, lower transparency, and higher uncertainty of high frequency markets all contribute to a sense that markets can be more fair for some than for others. How, exactly, to investigate this hypothesis is complicated because fairness is hard to define and even harder to measure. It could be easier to assess unfairness on an ex post basis as a likely manifestation is an unwillingness of individuals to participate in markets.[21] To the extent this happens, markets fail to provide risk sharing for individuals and access to risk capital for firms and entrepreneurs.

Many fairness issues concern features of the microstructure. Traders are not the same, so it makes sense that market also need not be the same. But does that mean that each market can choose exactly the microstructure it wants without regard to the overall effects on the market? For example, should exchanges be allowed to offer specialized order types targeted at high frequency traders if they result in advantaging their orders over those of other traders? Should exchanges be allowed to offer co-location for a fee, or must they provide the same access to every trader? Is it fair for exchanges to sell some traders information before it is available to others?

I think these questions are important and troubling. They also reveal interesting lacunae in the national market system. The 1975 Amendment to the Securities and Exchange Act of 1934 set out a framework of five principles to guide the development of a national market system. Only one mentions fairness, and that is in the context of competition between broker-dealers and exchanges ("Fair competition among broker-dealers, among exchanges, and between exchanges and other markets"). The notion of fairness with respect to different groups of traders was not on the radar screen.

All this suggests a need to do more policy-oriented research in microstructure. My own view is that answers to some of the questions posed above will require new regulations and changes in market practices. But to make this case more research is needed and, like everything else in today's markets, it needs to be done quickly.

[21] Guiso, Sapienza, and Zingales (2008) address the role of trust in the stock market and find that less trusting investors chose not to participate. Easley and O'Hara (2010) investigate the role of microstructure in ameliorating ambiguity in markets. They show how faced with ambiguity, traders also opt not to participate.

References

Angel, J., Harris, L., Spatt, C., 2011. Equity trading in the 21st century. Quarterly Journal of Finance 1 (1), 1–53.

Battalio, R., Corwin, S., Jennings, R., 2013. Can brokers have it all? On the relation between make-take fees and limit order execution quality. Unpublished working paper. University of Notre Dame.

Battalio, R., Shkilko, A., Van Ness, R., 2011. To pay or be paid? The impact of taker fees and order flow inducements on trading costs in US options markets. Unpublished working paper.

Berman, G., 2014. What drives the complexity and speed of our markets? US Securities and Exchange Commission Speeches, April 15. ⟨www.sec.gov/speeches⟩.

Biais, B., Foucault, T., Moinas, S., 2012. Equilibrium high-frequency trading. Unpublished working paper. University of Toulouse, Toulouse, France.

Biais, B., Wooley, P., 2011. High frequency trading. Unpublished working paper. University of Toulouse, Industrial Economics Institute, Toulouse, France.

Bloomfield, R., O'Hara, M., Saar, G., 2015. Hidden liquidity: some new light on dark trading. Journal of Finance. (forthcoming).

Brogaard, J., Hendershott, T.J., Riordan, R., 2013. High frequency trading and price discovery. Review of Financial Studies 27 (8), 2267–2306.

Brogaard, J., Garriott, C., Pomeranets, A., 2013. Is more high frequency trading better? Unpublished working paper. University of Washington.

Budish, E., Cramton, P., Shim, J., 2013. The high frequency trading arms race: frequent batch auctions as a market design response. Unpublished working paper. University of Chicago.

Carrion, A., 2013. Very fast money: high-frequency trading on Nasdaq. Journal of Financial Markets, 680–711.

Cespa, G., Vives, X., 2013. The welfare impact of high frequency. Unpublished working paper. IESE Business School.

Duffie, D., Garleanu, N., Pedersen, L., 2005. Over-the-counter markets. Econometrica 73, 1815–1847.

Easley, D., Kiefer, N.M., O'Hara, M., Paperman, J., 1996. Liquidity, information, and less-frequently traded stocks. Journal of Finance 51, 1405–1436.

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. Journal of Portfolio Management 37, 118–128.

Easley, D., Lopez de Prado, M., O'Hara, M., 2012. Flow toxicity and liquidity in a high frequency world. Review of Financial Studies 25, 1457–1493.

Easley, D., Lopez de Prado, M., O'Hara, M., 2013. Discerning information from trade data. Unpublished working paper. Cornell University.

Easley, D., O'Hara, M., 2010. Microstructure and ambiguity. Journal of Finance 65, 1817–1866.

Easley, D., O'Hara, Yang, L., 2015. Differential access to price information. Journal of Financial and Quantitative Analysis. (forthcoming).

Egginton, J.F., Van Ness, B., Van Ness, R., 2011. Quote stuffing. Unpublished working paper.

Foucault, T., Kadan, O., Kandel, E., 2005. Limit order book as a market for liquidity. Review of Financial Studies 18, 1171–1217.

Gloston, L., Milgrom, P., 1985. Bid, ask, and transaction prices in a securities market with heterogeneously informed traders. Journal of Financial Economics 14, 71–100.

Goldstein, M., Kumar, P., Graves, F.C., 2014. Computerized and high frequency trading. The Financial Review 49 (2), 177–202.

Guiso, L., Sapienza, P., Zingales, L., 2008. Trusting in stock markets. Journal of Finance 63 (6), 2257–2260.

Haldane, A, 2011. The race to zero. Bank of England speeches, given to the International Economic Association 16th World Congress, July 8.

Haldane, A., 2012. Financial arms races. Bank of England speeches, delivered at the Institute for New Economic Thinking. Berlin, Germany, April 14.

Hasbrouck, J., 2013. High frequency quoting: short-term volatility in bids and offers. Unpublished working paper. ⟨http://ssrn.com/abstract=2237499⟩.

Hasbrouck, J., Saar, G., 2013. Low-latency trading. Journal of Financial Markets, 646–679.

Hasbrouck, J., Sofianos, G., 1993. The trades of market makers: an empirical analysis of NYSE specialists. Journal of Finance 48 (5), 1565–1593.

Hendershott, T., Jones, C., Menkveld, A.J., 2013. Implementation shortfall with transitory price effects. In: Easley, D., Lopez de Prado, M., O'Hara, M. (Eds.), High Frequency Trading: New Realities for Trades, Markets, and Regulators, Risk Books, London, UK, pp. 185–206.

Ho, T., Stoll, H., 1981. Optimal dealer pricing under transactions and return uncertainty. Journal of Financial Economics 9 (1), 47–73.

Holden, C., Jacobsen, S., 2013. Liquidity measurement problems in fast, competitive markets: expensive and cheap solutions. Journal of Finance, 1747–1785.

Holden, C., Jacobsen, S., Subrahmanyam, A., 2014. The empirical analysis of liquidity. Unpublished working paper. University of Indiana.

Jarrow, R.A., Protter, P., 2012. A dysfunctional role of high frequency trading in electronic markets. International Journal of Theoretical and Applied Finance 15 (3), 2–15.

Jones, C., 2012. What do we know about high frequency trading? Unpublished working paper. Columbia University.

Kirilenko, A.A., Kyle, A.S., Samadi, M., Tuzun, T., 2011. The flash crash: the impact of high frequency trading on an electronic market. Unpublished working paper. Commodity Futures Trading Commission and University of Maryland, Washington, DC, and College Park, MD.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 53 (6), 1315–1335.

Laughlin, G., Aquirre, A., Grundfest, J., 2012. Information transmission between financial markets in Chicago and New York. Unpublished working paper. Stanford Law School, John M. Olin Program in Law and Economics, Stanford, CA.

Linton, O., O'Hara, M., Zigrand, J.P., 2013. The regulatory challenge of high frequency markets. In: Easley, D., Lopez de Prado, M., O'Hara, M. (Eds.), High Frequency Trading: New Realities for Trades, Markets, and Regulators, Risk Books, London, UK, pp. 207–230.

Madhavan, A., 2013. Exchange-traded funds, market structure, and the flash crash. Financial Analysts Journal 68, 20–35.

Malinova, K., Park, A., 2013. Do retail traders benefit from improvements in liquidity? Unpublished working paper. University of Toronto.

Menkveld, A.J., 2013. High frequency trading and the new market makers. Journal of Financial Markets, 712–740.

Menkveld, A.J., Yueshen, B., 2013. Anatomy of the flash crash. Unpublished working paper. University of Amsterdam.

O'Hara, M., 2011. What is a quote? Journal of Trading 5 (2), 10–26.

O'Hara, M., Saar, G., Zhang, Z., 2013. Relative tick size and the trading environment. Unpublished working paper. Cornell University.

O'Hara, M., Yao, C., Ye, M., 2014. What's not there: odd lots and market data. Journal of Finance. 69 (5), 2199–2236.

Pagnotti, E., Phillipon, T., 2013. Competing on speed. Unpublished working paper.

Securities and Exchange Commission, 2013. Data highlights 2013-05: the speed of equity markets. October 9. ⟨www.sec.gov⟩.

Securities and Exchange Commission, 2014. Data highlight 2014-02: equity market speed relative to order placement. March 19. ⟨www.sec.gov⟩.

Virtu Financial, Inc., 2014. Form S-1. Filed with the Securities and Exchange Commission. March 10.

Ye, M., Yao, C., 2014. Tick size constraints, market structure, and liquidity. Unpublished working paper. University of Illinois.

Ye, M., Yao, C., Jiading, G., 2013. The externalities of high-frequency trading. Unpublished working paper. University of Illinois.