High Frequency Market Microstructure

Maureen O'Hara*

April 2014

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. 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, the research we do 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.

* Johnson Graduate School of Management, Cornell University. My thanks to Bill Schwert and Ken French for suggesting this paper to me. I am very grateful to Ayan Bhattacharya, David Easley, Frank Hatheway, Joel Hasbrouck, David Meitz, 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.

Disclosure: Maureen O'Hara is Chairman of the Board of Directors of ITG, a global broker-dealer firm focusing on the needs of buy-side institutional clients.

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 new role played by microstructure. One might have expected that when things are fast the market structure becomes irrelevant – the opposite is actually the case. At very fast speeds, only the microstructure matters.

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. Exchange matching engines become the focal point of high frequency strategies, meaning that how the market is structured becomes very important. HF strategies can be quite complex, but so, too, now are the strategies that other traders elect, in part because they need to optimize in a market that contains HF players. And the exchanges as well act strategically, 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 is now different, and the data that emerges from the trading process is 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 of the more important aspects of this high frequency transformation. For finance researchers more generally, understanding how markets and trading have changed is important

for informing future research. By describing the current "lay of the land", I hope this paper can facilitate this understanding. For microstructure researchers, I believe these changes call for a new research agenda, one that recognizes that the learning models we used in the past are now deficient and that the empirical methods we traditionally employed may no longer be appropriate. Equally important, I believe that microstructure research has to provide more policy guidance, reflecting the problem that the new complexity of markets may confound even the best- intentioned regulators.

Some of this new agenda for high frequency market microstructure research is well underway, 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 fulsome reviews are Biais and Wooley [2011); ; Angel, Harris, and Spatt [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 we need to address in microstructure research.

This paper is organized as follows. The next section describes the new high frequency world, with a particular focus on how HFT affects the strategies of traders and markets. I discuss the varied behaviors and strategies of high frequency traders, and how trading for other non-high frequency traders has changed. I also discuss how HFT has affected the organization of trading, with particular attention given to the role of exchange pricing models, order priority rules, and the development of new trading platforms. In Section 3, 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, the research we do 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. Section 4 is a conclusion.

2. The New High Frequency World

It is not exactly clear when high frequency trading became a dominant force in markets, but certainly 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. My focus will primarily be on equity trading, but a similar transformation has taken place in other asset classes, with the trading of foreign exchange, futures, and options markets all now radically different.

A. High Frequency Traders

A natural starting point is with the 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. Indeed, the SEC struggled to define HFT, ultimately adopting a characteristics-based definition.[1] Certainly, all HFT activities have some things in common. HFT is done by computers, also referred to as silicon traders, it relies on extremely fast speeds, and it is strategy-based. But within HFT, there can be large differences even in these common traits.

[1] In its concept release on market structure, the SEC lists several characteristics commonly attributed to HFTs including "(1) the use of extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders; (2) use of co-location services and individual data feeds offered by exchanges and others to minimize network and other types of latencies; (3) very short time-frames for establishing and liquidating positions; (4) the submission of numerous orders that are cancelled shortly after submission; and (5) ending the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions over-night)." See SEC [2010].

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).[2] While the low latency group relies on co-location of servers within exchanges and fast, dedicated access to trading information, the ultra group augments this with enhancements like the Hibernian Express (the new undersea cable linking London markets and U.S. markets scheduled to come on-line in mid-2014), Perseus Telecom's new microwave network between London and Frankfurt (which reduces round trip latency to below 4.6 milliseconds from the 8.35 milliseconds using a fiber-optic network), and the development of new micro-chips capable of sending trades in 740 billionths of a second (or nanoseconds).[3] Hasbrouck and Saar [2013] using order level data from 2008 find that some traders in Nasdaq could respond to events such as changes in the limit order book in 2-3 milliseconds. Ye, Yao, and Jiading [2013] provide evidence of trading at even faster speeds. A particular high frequency trader's need for speed will depend upon the specific strategies they pursue.

All HFT is strategic because its goal is generally to be the "first in line" to trade. This is where microstructure comes to the fore because how to achieve this goal depends on the rules and structure of the market (i.e. its microstructure). At a minimum, this requires maximizing your trading strategy against a particular market's matching engine.[4] The matching engine

[2] For an excellent discussion of the distinctions between low latency, ultra-low latency, (and even ultra-ultra-low latency) HFT trading in foreign exchange markets see Olsen et al [2013].

[3] See "Lasers, microwave deployed in high speed trading arms race," Reuters, May 13, 2013, and "Wall Street's Need for Trading Speed: The Nanosecond Age", The Wall Street Journal, June 14, 2011, cited in Ye, Yao, and Jiading [2012].

[4] For a discussion of the Eurex matching engine see "Some insights into the details that matter for high frequency trading", Eurex, November 2013 at http://www.eurexchange.com/blob/exchange-en/4038-4046/238346/9/data/presentation_eurex_trading_systems_en.pdf

determines how orders are processed, and thus how trades and prices emerge. The matching engine also processes messages regarding the arrival, execution, and cancellation of orders.

Exchanges generally offer different connection speeds to their matching engines providing high frequency traders with the latencies that they need to implement trading strategies. Figure 1 illustrates the trading platform of the Tokyo Stock Exchange, which offers three different levels of connectivity services.[5] The Arrownet line is the standard service providing latencies in the range of several milliseconds. Faster still is their priority service which allows users to put devices at the data centers. Even faster connection is available if users co-locate trading devices at the TSE primary site. Of course, the Exchange charges higher fees for using these lower latency access paths to the matching engine.

Exchanges use different priority rules to handle orders. Price - time priority in which orders with the best price trade first, and among those with the same price the first order to arrive has priority, has been the most common rule in equity markets. Other priority rules do exist, however, such as price-size-time priority which favors those willing to trade larger sizes. Another priority structure often found in futures markets (and in some equity crossing networks) is pro-rata matching, in which all orders at a given price trade proportionately against an incoming order. How you trade will differ depending upon these order handling rules. Exchanges also have different pricing rules (these are discussed in more detail shortly), and HFTs optimize against these as well.

HFTs pursue a wide range 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 [2012], Carrion [2013]). HFT market

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

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 (i.e. covariances) in price ticks to "lift" liquidity between securities or markets. For example, if an upward price tick in one contract or security is generally followed by a similar upward price tick in another contract, an HF market maker will sell in the first market and buy in the second. This involves submitting an order at the ask in one market and at the bid in the other market.

In a market with maker-taker pricing, these orders will generally (but not always) be limit orders, meaning that the HF trader is supplying liquidity just as in more traditional market making. But unlike its traditional counterpart, the HF market maker is only on one side of the book in each market, and there is no commitment to provide liquidity continuously (see Menkveld [2012] and Virtu [2014] for empirical evidence on HF market making)[6]. This has led to concerns that HFT market making can induce market instability (see Kirilenko et al [2011], Easley et al [2012], and Madhavan [2013]). It remains the case, however, that the bulk of liquidity provision in many markets is provided by high frequency traders (see Brogaard [2011], and O'Hara, Saar, and Zhang [2013]).

Other HFT strategies employ more complex opportunistic algorithms to take advantage of profit opportunities (see Hagstromer and Norden [2013] for more discussion). Of particular importance are strategies designed to exploit predictable trading patterns of non-HFT traders. Some of these strategies are fairly straightforward, such as exploiting the deterministic behavior of simple algorithms like TWAP (time-weighted average pricing). Other strategies are more

[6] Virtu Financial, Inc., a leading high frequency market making firm, recently filed an S1 in advance of their proposed IPO. 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 page 100 of Virtu [2014]).

devious such as momentum ignition strategies designed to elicit predictable price patterns from orders submitted by momentum traders.

Some HFT strategies cross the line into unethical behavior.[7] 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 may yield immediate profits (the HFT can sell at the now higher price) or more circuitous returns (the HFT may profit by trading in a crossing network at the now higher mid-quote price). In either case, this predatory strategy is a form of "spoofing" and as such is forbidden under Dodd-Frank. Edderington, Van Ness, and Van Ness [2011] and Ye, Yao and Jiading [2013] (YYJ) examine another manipulative strategy called "quote stuffing". This strategy involves sending and instantly cancelling massive number of orders to the exchange with the designed purpose of slowing down trading for rival HFT firms..

There is now increasing awareness by both regulators and other market participants that HFT endeavors are best separated into "good" activities and "predatory" activities. And 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. These changes highlight why HFT has been so transformational – speed is only a small part of the story.

B. Non-High Frequency Traders (i.e. Everybody Else (EE))

In the previous sub-section, I noted that HFT is done by computers, relies on extremely fast speeds, and is strategy-based. What may not be fully appreciated is that this also describes

[7] See Biais and Wooley [2011] for an excellent discussion of such strategies. See also Jarrow and Protter [2012] for a model of high frequency manipulative strategies.

non-HFT trading. 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 sub one mili-second ranging down to 500 micro-seconds for a market order sent via DMA (direct market access). Such speeds were unheard of for even HFTs a few years ago, let alone for our "EE" trading! Of course, the speeds at the pure ultra-latency shops are even faster, with some claiming round trip latencies of sub 10 micro-seconds.[8] The bottom line is that trading is now very fast for everyone in the market.

It is also now 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 for both institutions and retail traders. Even without the complications introduced by HFT, trading is a challenging task in the fragmented market structure of current U.S. equity markets. Finding, and accessing, liquidity generally requires routing orders to multiple locations, all the while being cognizant of differing 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 deterministic trading patterns of unsophisticated traders and take advantage of them, it is little wonder that EE trading now relies on increasingly sophisticated trading algorithms.[9]

These trading algorithms run the gamut from simple trading rules like TWAP (designed to execute orders at the time-weighted average price over some interval) and VWAP (designed to

[8] The latency numbers cited here are for orders sent to the venue and then acknowledged back to the customer. If the order is then routed out of the broker/dealer to an exchange, there is additional time involved, raising latencies to around 4 to 5 milliseconds. My thanks to David Meitz and Jamie Selway for insights on this issue.
[9] Determining whether HFT can detect non-high frequency customer orders being traded by algorithms then becomes an important issue for algorithm designers. See Sofianos and Yang [2013] for empirical evidence on this issue for a specific set of algorithms used by a large broker-dealer firm.

get the volume-weighted average price) to more tactical strategies employing dynamic order strategies in both lit and dark markets. Table 1 gives a sample of the main algorithms currently offered by a large global broker/dealer firm to its buy-side institutional clients. These algorithms fall into general categories of dark aggregation (relying on crossing networks and other non-displayed order strategies), scheduled (chopping orders up deterministically as in TWAP), 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).

The diversity of these algorithms reflects the complex trading problem facing non-high frequency traders in current markets. Where to trade, how quickly to trade, whether to take liquidity via market orders or make liquidity via limit orders, or to trade at the mid-point in a crossing network or ATS, or to hide trading intentions via hidden orders on exchanges or in dark pools – these are but a few of many dimensions involved in formulating trading strategies for non-high frequency traders. An interesting feature of most trade algorithms is that they are rarely "pure" strategies – most algos, for example, both supply and demand liquidity, and algos typically transact across a variety of market venues.

This strategic trading by "everybody else" is one reason why dark trading has become so important. It also explains why trade sizes have fallen so dramatically. Figure 2 shows that the average trade size on U.S. equity markets is now just over 200 shares. Perhaps more significant is that the median trade size (at least on the Nasdaq) is now 100 shares, and odd lot trades are upwards of 20% of all trades (see O'Hara, Ye, and Yao [2013]). It is important to stress that the diminution of trade size is not just driven by market fragmentation. Futures markets are not fragmented, but, as Figure 3 shows, trade sizes are rapidly falling, with the average trade on

Treasury Bond Future now below 13 contracts. The average trade size on the WTI Crude Oil Future is even lower, falling to an average of 1.2 contracts per trade. 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 now converging to one contract at a time.

One might expect that retail traders fare poorly in this environment, but this misses the reality that retail trading is also now changed. In U.S. markets, a large fraction of retail trades are either directly internalized or delivered via purchased order flow agreements to large broker/dealers.[10] For example, Charles Schwab's order flow currently is "sold" to UBS, which because of best execution requirements must provide the prevailing best bid or ask for these orders.[11] 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.[12]

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 actually high frequency trading firms.[13] Trading costs of retail traders have been falling generally over the past 30 years (see Angel, Harris, and Spatt [2011]), but this decline seems to have

[10] Sophisticated retail traders often trade through firms such as Interactive Brokers, which provide state of the art trading tools and advanced trading algorithms to retail traders.

[11] Similarly, due to purchased order flow arrangements, TD Ameritrade's order flow is routed to Citadel.

[12] In particular, BCJ provide evidence that some large retail brokers sell all their order flow in purchased order arrangement, while others sell their market orders but rout their limit orders to the venues giving the highest liquidity rebates (in their data set Edge X was paying the highest amount).

[13] 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 O'Hara, Saar, and Zhang [2013] for discussion.

accelerated. Using data from the Toronto Stock Exchange, Malinova and Park [2013] provide empirical evidence that retail trading cost have fallen because of the presence of HFTs.[14]

C. 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 these microstructure issues: with HFT more than half of all trading volume, making its microstructure more attractive to high frequency traders entices needed volume (and liquidity) to the exchange; but becoming too HFT-friendly risks alienating EE traders (the institutional and retail traders) who may then choose to trade in specialized venues elsewhere.

This dilemma over HFT market design is only the latest chapter in a debate that pre-dates the high frequency era. As trading became increasingly electronic, the SEC struggled with how to craft a regulatory framework that would encourage innovation in markets while at the same time promote competition between markets. Reg ATS in 2000 and Reg NMS in 2007 set out a market framework in which new competitors (the alternative trading systems) were allowed to enter, and competition in this national market system was ensured by enhanced trade-through rules and explicit market linkage regulations. With the end of the "one-size-fits-all" model of exchange trading, new trading venues crafted microstructures to meet the particular needs of specific traders. And exchanges, faced with competition from all sides, responded by creating markets within markets, essentially setting up specialized microstructures to attract particular trading clienteles.

The end result is that trading is now 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.[15]

[14] Hendershott, Jones and Menkveld [2011] were among the first to show that algorithmic trading improved overall market quality in terms of improved liquidity and enhanced informativeness of quotes. Boehmer, Fong, and Wu [2013] confirm this more generally using data from 39 markets.

Exchanges and trading venues face intense competition to get the "right" order flow, avoiding if at all possible the toxic orders that disadvantage other traders. This competition is all the more complicated in the case of high frequency trading as, per our previous discussion, trading venues want to attract the "good" HFTs but not the "bad" HFTs. The key to doing so involves a variety of strategic decisions with respect to market design.

Certainly, one such strategic decision involves 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 trader who hit existing orders via market orders are said to "take" liquidity. Island ECN first 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 HFTs.[16]

But maker-taker is not the only way to structure a market. As Table 2 shows, 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 where you can trade as much as you want for a given monthly fee.[17] The taker-maker pricing strategy is particularly interesting, and it harkens back to the notion that

[15] BATS, now the third largest equity exchange in the US, illustrates how this new competition works. They initially began trading as an electronic order book (technically an ATS – or in their parlance, a better alternative trading system (BATS)). They attracted order flow by essentially giving away trading by cutting trading fees well below those of market competitors. Having quickly gotten market share, they then were able to retain it by having faster technology and a more flexible market structure. BATS then shifted from being an ATS to becoming an exchange. Their merger last year with Direct Edge (a trading venue with a similar evolution) gave them even greater scale. BATS is also a major trading venue in European markets.

[16] See Kearns, Kulesza, and Nevmyvaka [2010] and Menkveld [2013] for analysis of HFT profits.

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

certain types of 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.[18] Ye and Yao [2014] argue that taker-maker pricing also provides a way for non-HFT 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 may feature multiple trading platforms, a trading venue can attract some clienteles to one platform and different clienteles to another (or, as the case may be, the same clientele simply pursing different strategies on each platform). For example, the BATS exchange features four trading platforms – the BZX Exchange, BZY Exchange, EdgX and EdgA. The BZX and EdgX platforms feature maker-taker pricing, while the BZY and EdgA platforms feature taker-maker.[19] The BZX Exchange 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 is intended to enhance market quality by incentivizing liquidity provision at the current bid and offer.

Exchanges also use different order types to appeal to high frequency traders. For example, Direct Edge introduced "Hide not Slide" orders, a complex order type allowing particular orders in locked markets to be hidden rather than have to move to a (worse) price that unlocked the market. These orders then revert to regular limit orders when the market unlocks,

[18] 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 CBOE and ISE). Battalio, Shkilko, and Van Ness [2011] provide an interesting analysis of the effects of these option market pricing schemes on execution quality.

[19] EdgX and EdgA are both trading platforms that were on the Direct Edge Exchange. Direct Edge merged with BATS, and the combined entity retained all four platforms.

but with the advantage of being first in the queue.[20] BATS features a similar BATS-only-Post-only order, NYSE Arca calls its version a Post-no-preference Blind Order, and Nasdaq has a Price-to-comply order. The queue-jumping feature of these orders has elicited complaints that these orders unfairly disadvantage other traders in the market. Another controversial order type is the PL Select Order introduced by NYSE/ Arca. This type of conditional limit order allowed the submitter to limit the contra side for execution, giving high frequency traders the ability to post orders that will only transact against the small (non-toxic) orders normally sent by small traders.

Trading venues also compete to attract HFT order flow via features such as 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.[21] The NYSE's proposed venture with Anova Technology using laser-millimeter wave technology may be even faster. For HFT's seeking the fastest way to trade, these technological innovations are key to deciding where to trade; for exchanges and markets, providing these innovations is key to their competitiveness (and survival). Whether the market overall (or society in general) is enhanced by such an arms race in technology is debatable (see Brogaard, Garriott, and Pomeranets [2013], Cespa and Vives [2013], Haldane [2012], Pagnotta and Philippon [2011], Bias, Foucault, and Moinas [2013], and Budish, Cramton, and Shim [2013]). What is not in question is how expensive these technologies are – Laughlin et al [2012] estimate that achieving a 3-millisecond decrease in communication time between Chicago and New York markets cost a staggering $500 million.

[20] See "How 'Hide Not Slide' Orders Work", Wall Street Journal, Sept. 18, 2012.
[21] For more discussion, see "High-Speed Stock Traders Turn to Laser Beams", Wall Street Journal, Feb. 11, 2014, and "Lasers, microwave deployed in high-speed trading arms race", Reuters, May 1, 2013 at www.reuters.com/assets/print?aid=USBRE9400L920130501.

Conversely, markets designed specifically to limit the involvement of HFTs are yet another dimension of strategic competition. In particular, the IEX in the U.S and Aequitas in Canada have designed microstructures to protect EE traders from HFTs. The IEX launched on October 25, 2013 as a broker-backed dark pool featuring price-broker-time priority. Thus, orders from an agency broker would have higher priority than an order coming from a high frequency trader. The IEX charges both sides of the transaction, so maker and taker designations are irrelevant. Orders in this market are all also "slowed down" with orders delayed by a randomized 10 or 15 microseconds. This feature is designed to limit the advantages of speed and so negate any advantage to the high frequency traders.

The Aequitas market in Canada 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 instead the market design relies on priority given to market makers to incentivize liquidity provision. Only retail and institutional traders are permitted to take liquidity in the market, in line with its goals to be a better venue for the "EE" traders.

A within-market approach to accomplish a similar goal for retail traders is the NYSE's Retail Liquidity Program (RLP). As noted earlier, little retail flow comes to the exchange due to diversion by purchased order arrangements, internalization more generally, or algorithms that route orders to other locales (such as crossing networks) before sending them to the exchange. The RLP is intended to "promote price improvement for individual investors on retail order flow for NYSE and NYSE MKT securities". 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).[22] Executed trades are required to receive price improvement relative to the BBO of at least 0.001, with RLPs quoting in increments of .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 can succeed in luring retail trade away from other trading venues. One thing it has attracted is competition from competing venues - BATS has set up a similar retail-focused program called RPI.[23]

The new high frequency world is thus both complex and constantly evolving. New technology and greater speed lead to new strategies, which lead to new methods of trading, and, in turn, to new market designs. But 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 to the fundamental properties of market data such as buys and sells, quotes and prices. These changes, in my view, are equally important to understand because they challenge the ways researchers have interpreted market data and analyzed market behavior and performance. In the next section, I examine these changes in more detail and make the case for a new agenda for microstructure research.

3. Microstructure Research: What is (or should be) Different?

[22] The program also allows retail traders, depending upon the specific order type, to interact with other liquidity coming from floor brokers and NYSE member firms. For details see "Retail Liquidity Program" at www.nyx.com

[23] Details of BATS Retail Price Improvement program can be found at http://cdn.batstrading.com/resources/regulation/rule_book/BYX_Rulebook.pdf.

As I argued in the previous section, high frequency trading has transformed markets, and in the process it has ushered in a new "golden age" for microstructure researchers. With markets and trading now radically different, there are myriad questions attracting the attention of researchers. These issues run the gamut from the particular – how do specific trading strategies affect market performance, to the more general – how has market quality fared in this new environment, to the more conceptual - how should markets be designed and what activities should regulation allow? Yet, while acknowledging the importance of this research, I believe that the high frequency world has also fundamentally altered some of the basic constructs underlying microstructure research. Consequently, in this section, I turn our attention to these more basic issues, with a goal of setting out some fundamental issues I believe are no longer well captured by our existing models and approaches.

A. Information in a high frequency world – what should we be learning?

Microstructure models are learning models.[24] In their canonical form, microstructure models rely on a basic story: some traders have private information; they trade on it; other traders see market data; they learn from it; market prices adjust to efficient levels that reflect the new information. Microstructure enters by influencing the types of market information that traders see and the ease with which they can learn from it. In this learning 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 hence "the market") also learn from other 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.

[24] This, of course, refers to the information-based microstructure models. There are also inventory models and search-based models in microstructure that generally eschew information issues.

In the high frequency world, the basic story remains the same: traders still have to trade to profit from information, and other traders will still try to learn what they know from watching market data. But some things are now very different. Traders are silicon, not human. Market data is 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 here 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 on the asset. 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, milliseconds, microseconds, perhaps even nanoseconds. Over these intervals, information may not just be asset-related but order-related as well. Haldane [2011] 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." [25]

This notion of speed being synonymous with informed trading is surely not the complete story, but it does speak to the complexities of information in the high frequency age. Informed trading is now multi-dimensional 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 are now selling

[25] See Haldane, A, [2011], page 4.

access to public information signals seconds (or even milliseconds) before they are seen by other traders. This effectively turns public information into private information, and corresponds, albeit for only a very short period, into the classic information-based trading of standard microstructure models. But HFTs can also turn speed into information via co-location and other technologies that allow them to process market data (such as prices and order book information) before everyone else.[26] If this information allows them to predict market movements better than other traders, then they, too, are clearly informed traders.

Indeed, even large traders who know nothing special about the asset's value can be lethal to market makers simply because they know more about the nature of their own trading plans. Trade imbalances are problematic for market makers because the market maker is always on the other side – buying if the traders are selling and selling if they are buying. Trading that is too heavily skewed to buys or sells is thus "toxic" and it can lead market makers to withdraw from trading as their inventory or short positions reach pre-set parameters (recall that the market makers are also simply algorithms so risk management is programmed in via limits on positions).[27] Over the short time intervals of interest to market makers, even these classically uninformed traders are informed traders in the new high frequency world.

From a research perspective, these expanded definitions of informed trading are worrisome. Now, it is not clear what is driving the adjustment of prices or, more to the point,

[26] Exchanges sell high speed data feeds as yet another strategy to entice high frequency traders to their markets. In the U.S., all trades and quotes must be reported to the consolidate tape at the same time they are sent out by the exchanges to their proprietary feeds, but in Europe there is not such tape. Easley, O'Hara and Yang [2012] show that this enhanced access to price data induces an adverse selection problem that reduces liquidity and increases the cost of capital for the economy. Biais, Foucault and Moinas [2013] demonstrate a similar effect with respect to fast information overall, and they argue that this gives rise to excessive investment in fast trading technologies.

[27] Indeed, Virtu [2014] stresses that this behavior is key to their risk management approach. They note "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 page 2 and page 109 of their prospectus).

where they are going. Analyses of market efficiency seem to suggest that markets generally remain informationally efficient, which should allay at least some concerns for asset pricing researchers. But episodic instability is also now characteristic of markets, driven perhaps by the desires of the "informed" high frequency market makers fleeing when they suspect other "more informed" traders are present. Markets also appear to be more tightly inter-connected, sewn together by market making/ statistical arbitrage that operates across rather than just within markets. These characteristics suggest that liquidity factors may play an increased role in asset pricing. What these liquidity factors are capturing, however, and how to even measure them, is problematic.

To understand these asset pricing issues, I think we need to understand better the new role (and definitions) of adverse selection, information, and liquidity at a microstructure level. The artificial divergence in microstructure models between those focusing on information issues and those focusing on inventory issues is now 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, is a proxy for competitive pricing in markets) may not be accurate given that pre-set risk limits induce non-participation by silicon traders providing liquidity.[28] The need for new and better microstructure models seems clear.

B. Market data - what are we looking at?

How to find the "informed traders" has always been a fundamental issue in microstructure models, and it speaks to the issue of learning from market data. The notion that

[28] One might have thought that "someone" overseeing the market making algorithm steps in and changes the program as daily conditions change, but this is not the case. Algorithms are written, back tested, and employed. If the process is not working well then the program is pulled and a new program developed. Indeed, testing of new trading algorithms is common in markets on non-live machines. It is alleged that a major problem in the Knight trading debacle was the failure to realize that the market making algorithm was now live in the actual market as opposed to being on the testing server.

informed traders leave "footprints" in markets is well established, and it is the reason why microstructure models attach such significance to trade data. Every trade has a buyer and a seller, but in microstructure we have been interested in the "active side" because of its signal value to the underlying information.[29] Buys were thus signals of good news; sells were signals of bad news. 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 complicates drawing inferences from market data in myriad ways. A fundamental problem is that because of algorithmic trading it is orders, and not trades, that reflect a trader's intentions. Algos chop a "parent order" into child orders, and it is these child orders (or some portion of them) that ultimately turn into actual trades. Unfortunately, neither the market (nor the researcher) can see these parent orders, and these child orders may have very different properties than are envisioned in microstructure models. For example, because the child orders are not independent, trades are not independent either, and sequences of trades now become informative (it is these patterns in trading that HFT often try to exploit).

Dynamic trading strategies also mean that these orders need not result in the simple buy and sell trades of times past. As Hasbrouck and Saar [2011] first pointed out, technology allows orders to be submitted (and cancelled) instantaneously, and dynamic strategies use this functionality to implement complex trading strategies. Because of maker/taker pricing, algorithms rely more on limit orders to reduce the transactions costs of trading. Algorithms also make extensive use of midpoint orders, a type of limit order that adjusts the limit order price to

[29] 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. Sequential trade models such as Glosten and Milgrom [1985] and the Kyle [1985) model both use buys and sells as the inputs to the market maker's pricing problem. Buy-sell data is also the main input data in PIN models (see Easley, O'Hara, [1996]), and it plays a role in models explaining liquidity linkages across markets (see Holden, Jacobsen, and Subrahmanyam [2014],)

the moving mid-point price, and they also take advantage of trading opportunities at the midpoint in crossing networks. Sophisticated traders only cross the spread when it is absolutely necessary.

To see why this matters for interpreting market data, consider a parent order to buy 5000 shares. Whereas in the old days of specialist trading this order would have executed 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 cancelled and updated as trading progresses. Because these are limit orders, any executing order will actually be the "passive" side of the trade – so this 5000 share buy order will show up in the data as many small sell trades!

Does this actually happen? To investigate these order strategy effects, I looked at execution data from ITG (a large broker/dealer firm) for a sample of equity executions in their standard VWAP algorithm in the year 2013.[30] The particular 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.[31] The VWAP algorithm will operate differently depending upon factors such as the size of the order and the customer's preference over how quickly to execute the order, so the data are broken out by participation rate (i.e. the order size as a percent of volume) and by order size as a percentage of the total volume executed for the day.[32]

[30] I thank Jeff Bacidore, Wenjie Xu, Cindy Yang, and Lin Jiang for providing the data and technical analysis.
[31] So, for example, an 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.
[32] The calculations are based dollar-value weighted averages. Share weighted averages yield similar results.

Table 3 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, meaning that 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 will show up in the data at least two-thirds of the time as "sell" orders – and including the midpoint orders this could be as high as 87%. Less than one in eight executed trades actually cross the spread and thus are the classic "buy" trades of microstructure models.

Trade size and intensity clearly affect these numbers. For small orders, 10.62% of executions were aggressive, and this fraction of aggressive trades gradually increases with order size until you reach very large trade sizes when it accelerates. For parent orders as large as 25% (50%) of the day's dollar weighted volume, aggressive executions were still less than a quarter (half) of executed trades. For massive orders, aggressive trading can exceed passive, but these trades are a miniscule fraction of the overall sample (.0005% of the total sample). Indeed, even orders above 10% of the day's volume are very rare (.06% of the sample).

The 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 executions. Orders trading faster than this are more aggressive, but are also exceeding rare.

That trading intentions and executed trades may 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-CFTC staff report identified the causal factor as a "large trader" submitting a large sell order at approximately 2:00 pm which then caused the market to fall precipitously. However, 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 actually involved large numbers of limit sell trades, meaning that this trader was actually providing liquidity to the market, rather than taking it!

A more fundamental issue is whether we can actually link "buy" and "sell" trades with underlying information. Easley, Lopez de Prado, and O'Hara [2012] argue that the active side of the trade is now more related to one's willingness to cross the spread than it is to information-based trading. In general, if signed orders are informative, then we would expect a greater order imbalance to have a greater effect on prices. But if 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 Figure 4 which shows the relationship between the Hi-Lo Price and the signed trade imbalance over ten-minute trading intervals for three stocks: Apple (APPL), Intuitive Surgical, Inc. (ISRG), and NX Stage Medical, Inc., (NXTM).[33] The data are from the Nasdaq HF data base and are for all trading on Nasdaq for these stocks in October 2010. The Figure shows virtually no relationships between order imbalance and the high-low price, consistent with signed order imbalance actually reflecting trading patterns of less sophisticated uninformed traders.

[33] The HF 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/sell indicators. I am grateful to Mao Ye for his help with this analysis.

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 thus are given either the best bid or the best offer. Informed traders (who are either 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. Bloomfield, O'Hara, andSaar [2013] 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.

But if we cannot learn from buys and sells, what should be looking at to infer underlying information? The HF world gives some clues in that HFT algos 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 I believe much more research is needed to ascertain what can be learned from this market data. Hasbrouck and Saar [2013] is a good example of this new research in that they highlight the role played by runs and sequences of high frequency trades in affecting market behavior and quality.

Even more important is to recognize that this data has to be looked at across markets, and not just within individual markets. High frequency algorithms operate across markets, using the power of technology to predict price movements based on the behavior of correlated assets.[34] But if order books are linked, then so, too, must be order flows and price behavior. Theoretically modelling such inter-relationships seems a daunting task, so empirical analyses focusing on the predictive power of market variables both within and across markets may be a good place to

[34] See Almgren [2013] for discussion and analysis of across-market HFT activity in fixed income markets.

start. Certainly, understanding the changing nature of market data is an important direction for future research.

C. Analyzing data – what should we be doing?

In some ways, the high frequency era is the best of times for empirical researchers. With trading electronic and computerized, there is a wealth of trading data, and new data sets are becoming ever more available. But this new treasure trove of data 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, we need some new tools in the microstructure tool box.

Consider, for example, the issues connected with the consolidated tape. In the U.S., all equity trades must be reported to the consolidated tape on a real time basis. Yet, the seeming precision of this statement is illusory. Odd lots, for example, were not reported to the tape, an omission largely explained by historical conventions. As O'Hara, Yao, and Ye [2012] demonstrate, with the rise of algorithmic trading odd lots play a very different role in the high frequency world, with some stocks having 50% or more of trades execute in odd lots (the average across all stocks was greater than 20%). Perhaps more important, these authors show that odd lots have high information content, consistent with informed traders using odd lots to hide their trades from the market. The SEC has recently changed course, now requiring odd lot reporting to the tape as of December 2013. But a variety of other microstructure data, such as Rule 602 (trade execution quality) statistics still do not include odd lot data, and of course historical data remain incomplete. Such missing data problems are a natural concern to researchers.

The consolidated tape has another problem – the data may be out of order. There are now 17 lit equity markets in the United States and 50 or more other trading venues. Each of these is reporting trades to the tape, but at differing latencies. The time stamps on the tape reflect when the trade report is received, so there can be sorting errors in trade occurrence across markets. There are also timing difficulties introduced by withdrawn quotes and cancelled orders, which are now integral to the way high frequency trading strategies operate. For some research purposes, this is not a problem. But for others, such as using the current quote to assign buy and sell directions, it is a show-stopper.

An equally important problem is that the "true" state of the market may not be visible to researchers using the standard TAQ database (MTAQ) which time-stamps data only to the nearest second. Holden and Jacobsen [2013] provide disturbing evidence that using MTAQ instead of DTAQ, the more expensive daily TAQ database time-stamped in milliseconds, results in massive errors in computing percent quoted spreads, effective spreads, realized spreads, and even findings of negative spreads. [35] Having performed extensive analyses to understand why these errors arise, Holden and Jacobsen propose ways to address these difficulties for researchers using MTAQ in empirical work such as interpolating time stamps and adjusting the data for withdrawn quotes.

The high frequency world also poses new challenges for empirical analyses using quote data. Quotes play an important role in microstructure because they traditionally have 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

[35] In particular, these authors note on page 3"For MTAQ, when compared to DTAQ, 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".

viewed as the current unconditional expected price. Quotes derive from actual orders on the book, but in high frequency markets the vast majority (indeed, by some estimates 98%+) of these orders are actually cancelled. Cancellations, 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 on the market such as a decrease in quote informational content, an increase in execution risk for traders, and a reduction in the reliability of the mid-point as a reference price for crossing networks. He also 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.[36]

The use of 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. Their VPIN (volume synchronized probability of informed trade) measure uses volume buckets and trade imbalances to estimate order flow toxicity. Using a volume clock reduces the bias in empirical analysis arising from irregularly-spaced data, a feature that is surely a characteristic of high frequency markets.

[36] Hasbrouck also suggests an alternative to the Holden – Jacobsen approach for interpolating time stamps for MTAQ data. His approach uses a randomized procedure to assign time stamps.

These market dynamics also require care in the application of existing empirical techniques – some of our 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 5 minutes later), for example, are now often negative. Such spreads have traditionally 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 may simply reflect that in HF settings, 5 minutes is a "lifetime", and so is not a meaningful time frame in which to evaluate trading. Perhaps 5 seconds or 15 seconds is a better horizon – or perhaps the realized spread is just not a useful concept any more.

Similarly, constructs such as permanent and transitory price effects are suspect, victims of the problem discussed earlier of what time frame actually constitutes transitory (milliseconds? seconds?) or permanent (10 minutes ?, hourly?, daily?)[37]. Estimations using numbers of trades (such as PIN estimation) are now 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, latency issues across venues leading to 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 new high frequency world. Data sets cannot keep up with the high frequency world because HFT keeps evolving. Replacing monthly TAQ (MTAQ) with daily TAQ (DTAQ) as suggested by Holden et al [2013] will help researchers, but DTAQ

[37] 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.

with its millisecond time stamps is already being challenged by trading that is 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 issues in accessing the market) may mean that it tells you little about the price at which you can actually trade.

This suggests to me that the changing nature of markets requires researchers to develop new tools for empirical analysis. In my own work, my co-authors and I have been working on new empirical measures of toxicity (VPIN), as well as new empirical approaches to classify trading activity. But there are myriad issues that need to be addressed, including fundamental questions such as do we need to actually analyze all of this data or can we instead use some sort of optimal sampling approach, or find simple nearly-sufficient statistics? The answer to these questions, of course, depends upon what it is you want to know – and in the high frequency world there is no shortage of things we do not yet understand.

4. Conclusions

This paper has analyzed the changes wrought by high frequency trading. I have argued that HFT is not just about speed, but instead reflects a fundamental change in how traders trade and how markets operate. These changes, in turn, have important implications for microstructure research, and I have suggested a variety of theoretical and empirical issues we need to address in high frequency market microstructure research.

What I have not yet discussed are the many policy and regulatory issues that would greatly benefit from increased research at the microstructure level. Regulation trails practice, and regulators around the world have struggled to catch up to the high frequency markets they

oversee. Some issues, in my view, have received far too little regulatory attention. These include issues of differential access to information, market linkage rules such as trade-at or trade-through, and potentially unfair order types or access protocols. Other regulatory efforts, such as the SEC's new MIDAS system for trade surveillance or the SEC's Naked Access Rule, appear to be well thought out and appropriate for high frequency markets. Yet other proposals seem to me ill-conceived and out-of-touch with the new market realities (transactions taxes or proposals to cap messages, or cancellations, or the use of algorithms, would fall into this category). But to make this case we need to do more research – and, like everything else in today's markets, we need to do it quickly.

References

Almgren, R., 2013, Execution Strategies in Fixed Income Markets, in High Frequency Trading; New Realities for Trades, Markets and Regulators, Easley, D., M. Lopez de Prado, and M. O'Hara (editors), Risk Books (London: 2013).

Angel, J., L. Harris, and C. Spatt, 2011, Trading the 21st century, Quarterly Journal of Finance, 1 (1), 1-53.

Baruch, S., Glosten, L.R., 2013, Flickering quotes, Working paper, Columbia University.

Battalio, R., Corwin, S. and R. Jennings, 2013, Can Brokers have it all? On the Relation between Mak-take fess and Limit Order Execution Quality, Working Paper.

Battalio, R., Shkilko, A., and R. Van Ness, 2011, To Pay or Be Paid? The Impact of Taker Fees and Order Flow Inducements on Trading Costs in U.S. Options Markets, Working Paper.

Biais, B., Foucault, T., and Moinas, S., 2012. Equilibrium high-frequency trading. Working paper. University of Toulouse.

Biais, B. and P. Wooley, 2011, High Frequency Trading, Working Paper, Toulouse University, IDEI.

Bloomfield, R., O'Hara, M. and G. Saar, 2013, Hidden Liquidity: Some New Light on Dark Trading, Working Paper.

Boehmer E., Fong K. Y. L., Wu J., "International evidence on algorithmic trading", Working Paper.

Brogaard, J., Hendershott, T.J., Riordan, R., 2013. High-frequency trading and price discovery. Review of Financial Studies, *forthcoming*.

Brogaard, Jonathan, Garriott, C. and A. Pomranets, Is more high-frequency trading better?, Working Paper, Nov. 2012.

Brogaard, Jonathan, T. Hendershott, and Ryan Riordan, High Frequency Trading and Price Discovery, Working paper available at: http://ssrn.com/abstract=1928510

Brogaard J., Hendershott T., Hunt S., Latza T., Pedace L., Ysusi C., "High-Frequency Trading and the Execution Costs of Institutional Investors", FSA Occasional papers in Financial Regulation, January 2013

Burdish, E., Cramton, P., and J. Shim, The High Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response, Working paper.

Carrion, A., 2013. Very fast money: high-frequency trading on NASDAQ. Journal of Financial Markets.

Cespa, G. and X. Vives, The Welfare Impact of High Frequency, Working Paper, April 2013.

Chakrabarty, B., Jain, P K., Shkilko, A., and K. Sokolov, 2013, Quote Intensity and Market Quality: Effects of the SEC Naked Access Ban, at http://ssrn.com/abstract=2328231

Cvitanic J., and A.Kirilenko, "High Frequency Traders and Asset Prices", Working Paper, 2010

Easley, D., M. M. Lopez de Prado, and M. O'Hara. (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., M. Lopez de Prado, and M. O'Hara. (2012). Flow Toxicity and Liquidity in a High Frequency World. Review of Financial Studies 25:1457-1493.

Easley, D., M. Lopez de Prado, and M. O'Hara. (2012), Bulk Classification of Trading Activity, Working paper.

Easley, D., M. Lopez de Prado, and M. O'Hara. (2013), Optimal Execution Horizon, Mathematical Finance, *forthcoming*.

Foucault, T., J. Hombert, and I. Rosu (2013). News Trading and Speed. *Working Paper*.

Gai, J., Yao, C., and Ye, M., 2012. The externalities of high-frequency trading. Working paper. University of Illinois.

Goldstein, M., P. Kumar, and F. C. Graves, (2014), Computerized and High Frequency Trading, The Financial Review, 49 (2).

Hagströmer, B., Nordén, L.L., 2013. The diversity of high-frequency traders, Journal of Financial Markets.

Haldane, A., 2012, Financial Arms Races, Bank of England Speeches, 14 April 2012.

Haldane, A, 2011, "The Race to Zero", Bank of England Speeches, Speech given to the International Economic Association Sixteenth World Congress, July 2011.

Hasbrouck, J., 2013, High Frequency Quoting: Short-term Volatility in Bids and Offers, Working paper available at : http://ssrn.com/abstract=2237499.

Hasbrouck, J. and G. Saar (2011), "Technology and Liquidity Provision: The Blurring of Traditional Definitions," *Journal of Financial Markets*.

Hasbrouck, J., and G. Saar, 2013. Low-Latency Trading. Journal of Financial Markets.

Hasbrouck, J., and G. Sofianos, 1993, The Trades of Market Makers: An Empirical Analysis of NYSE Specialists, Journal of Finance, (48) 5.

Hendershott, T., C. Jones, and A.J. Menkveld. (2011). Does Algorithmic Trading Increase Liquidity? Journal of Finance 66:1-33.

Hendershott, T., C. Jones, and A.J. Menkveld. (2013), Implementation Shortfall with Transitory Price Effects, in High Frequency Trading; New Realities for Trades, Markets and Regulators, Easley, D., M. Lopez de Prado, and M. O'Hara (editors), Risk Books (London: 2013).

Hendershott, T., and A.J. Menkveld. (2011). Price Pressures. Working Paper.

Hendershott, T., and R. Riordan. (2012). Algorithmic Trading and the Market for Liquidity. Journal of Financial and Quantitative Analysis, *forthcoming.*

Hoffmann, P., 2013. A dynamic limit order market with fast and slow traders. Working paper. European Central Bank.

Holden, C. and S. Jacobsen, 2013, Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions, Journal of Finance, *forthcoming*.

Holden, C., S. Jacobsen, and A. Subrahmanyam, 2014, The Empirical Analysis of Liquidity, Working Paper.

Jarrow R. A., and P. Protter, "A Dysfunctional Role of High Frequency Trading in Electronic Markets", International Journal of Theoretical and Applied Finance, Vol. 15, No. 3, 2012

Jones, C., 2012, What do we know about high frequency trading?, Working paper

Jovanovic, B., Menkveld, A.J., 2010. Middlemen in limit-order markets. Working paper. New York University.

Kearns, M., Kulesza, A. and Y. Nevmyvaka, Empirical Limitations on High Frequency Profitability, Working Paper, 2010.

Kirilenko, A.A., Kyle, A.S., Samadi, M., and T. Tuzun, 2011. The flash crash: the impact of high frequency trading on an electronic market, Working paper. CFTC and University of Maryland.

Laughlin, G., A. Aquirre, and J. Grundfest, 2012, Information Transmission between Financial Markets in Chicago and New York, Stanford Law and Economics Working Paper.

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

Malinova, K. and A. Park, Do retail traders benefit from improvements in liquidity? Working paper, Nov. 2013.

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

O'Hara, M., 2011, What is a quote? Journal of Trading.

O'Hara, M., G. Saar, and Z. Zhang, 2013, Relative Tick Size and the Trading Environment, Working paper.

O'Hara, M. and M. Yao, 2011, Is Market Fragmentation Harming Market Quality?, Journal of Financial Economics.

Riordan R., and A. Storkenmaier, "Latency, Liquidity and Price Discovery", Journal of Financial Markets, 15(4), November 2012.

Sofianios, G., and J. Xiang, 2013, Do Algorithmic Executions Leak Information?" in High Frequency Trading: New Realities for Traders, Markets and Regulators, edited by Easley, D., M. Lopez de Prado, and M. O'Hara, Risk Books (London).

U.S. Commodities Futures Trading Commission and the U.S. Securities and Exchange Commission, 2010. Preliminary findings regarding the market events of May 6, 2010.

U.S. Securities and Exchange Commission, 2010, Concept release on equity market structure 34-61358.

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

Ye, M. and C. Yao, 2014, "Tick Size Constraints, Market Structure, and Liquidity, Working paper available at : http://ssrn.com/abstract=2359000

Ye M., Yao C., Jiading G., "The Externalities of High-Frequency Trading", Working Paper 2013

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.

General Type	Description	Uses
Opportunistic	Posit Marketplace	Access to dark liquidity in Posit and other dark venues.
	Raider	Operates strategically across both dark and lit markets to capture liquidity. It does not display in lit markets
	Float	Seeks to earn the spread by actively posting on the passive side
	Pounce	Opportunistic, liquidity-seeking, finding posted and reserve liquidity and employing pegged orders to wait for liquidity in illiquid stocks
	Flex	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	Optimizing 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 HFTs.

Source: ITG Alogrithms at http://www.itg.com/marketing/ITG_Algo_ExecutionStrategies_Guide_20130701.pdf

Table 2 Market Pricing Models

This table lists the main types of market pricing models and a sample of markets using these models.

Structure	**Venue**
Both sides pay	Bovespa, ICE, IEX, Deutsche Borse, HkEx, ITG, LX, Level ATS, Tokyo Stock Exchange
Maker – Taker	BATS BZX, EdgX, Nasdaq, NYSE, Arca, Chi-X, Lava ATS
Taker- Maker	Nasdaq BX, BATS BZY, EdgA, NYSE RLP, BOX
Subscription	Aquis

Source: Market web sites

Table 3. VWAP Execution Data

This table gives data from an VWAP algorithm executed for ITG buy-side client market orders. All parent orders are market orders, and are fully filled. All parent orders have at least 100 shares. Locked quotes are eliminated in Fill Aggressiveness and Trade statistic calculations. Percentages given are dollar-weighted. The upper panel gives data split by the participation rate (amount as a percentage of volume) of the order. The lower panel gives data split by the order size relative to that day's total volume. The sample period is 2013.

Participation Rate Bucket

Fill Aggressiveness			Dollar-Weighted		
Participation Rate	# Parent Orders	# Trades	Passive	Aggressive	Midpoint
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%

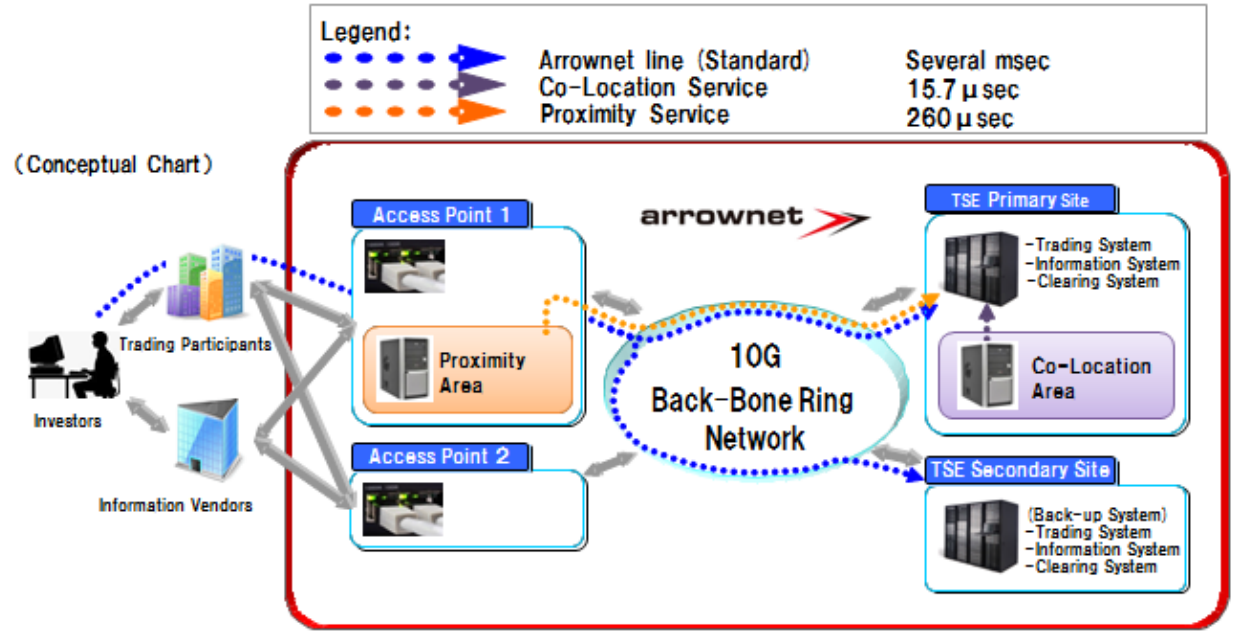
Order Size Bucket

Fill Aggressiveness			Dollar-Weighted		
Order Size	# Parent Orders	# Trades	Passive	Aggressive	Midpoint
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%

Source: ITG data

Figure 1. Connectivity on a Major Stock Exchange

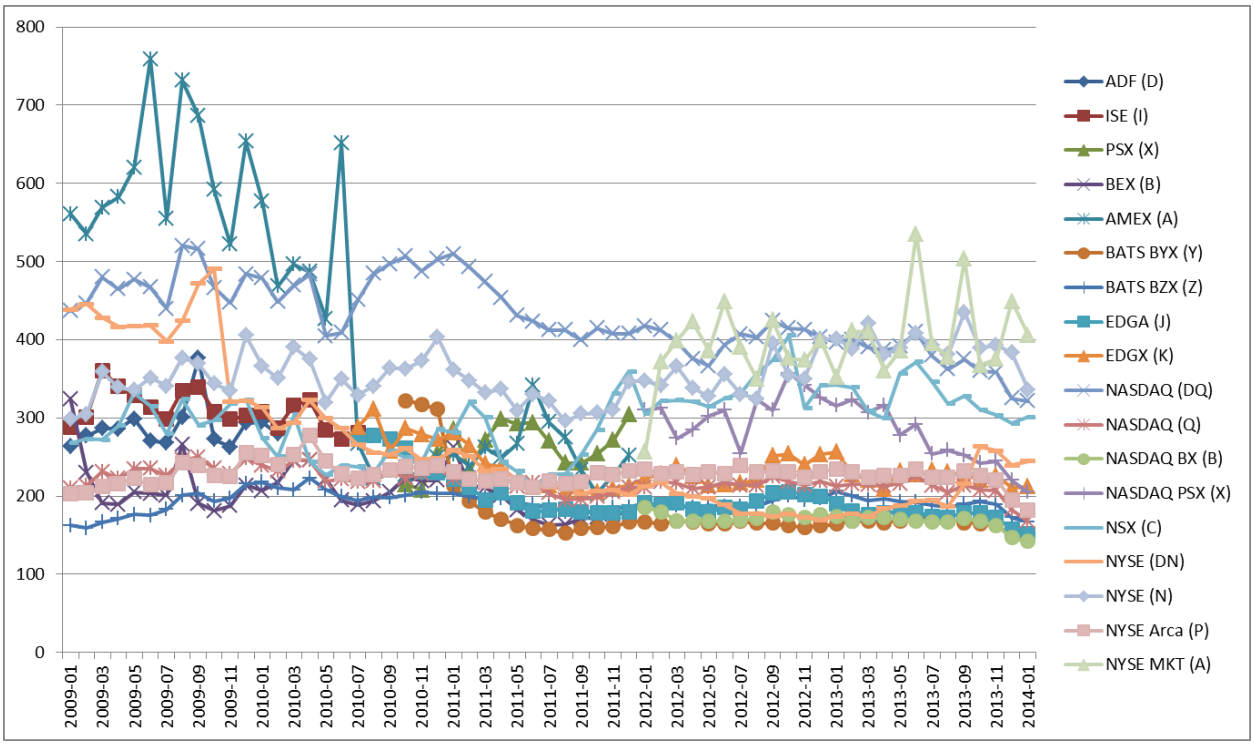
This figure shows the three connectivity options on the Tokyo Stock Exchange. Orders come in from traders and are routed to the exchange for execution. The standard connection is provided by the Arrownet line which provides latency measured in milliseconds (one-thousandth of a second). The proximity service allows for devices to be attached directly to the access point, thereby reducing latency to 260 microseconds (one-millionth of a second). The co-location option allows devices to be placed in the TSE's primary site, reducing latency to 15.7 microseconds.



Source: Tokyo Stock Exchange Web Site, http://www.tse.or.jp/english/system/connectivity/index.html

Figure 2. Average U.S. Trade Size in Equity Markets.

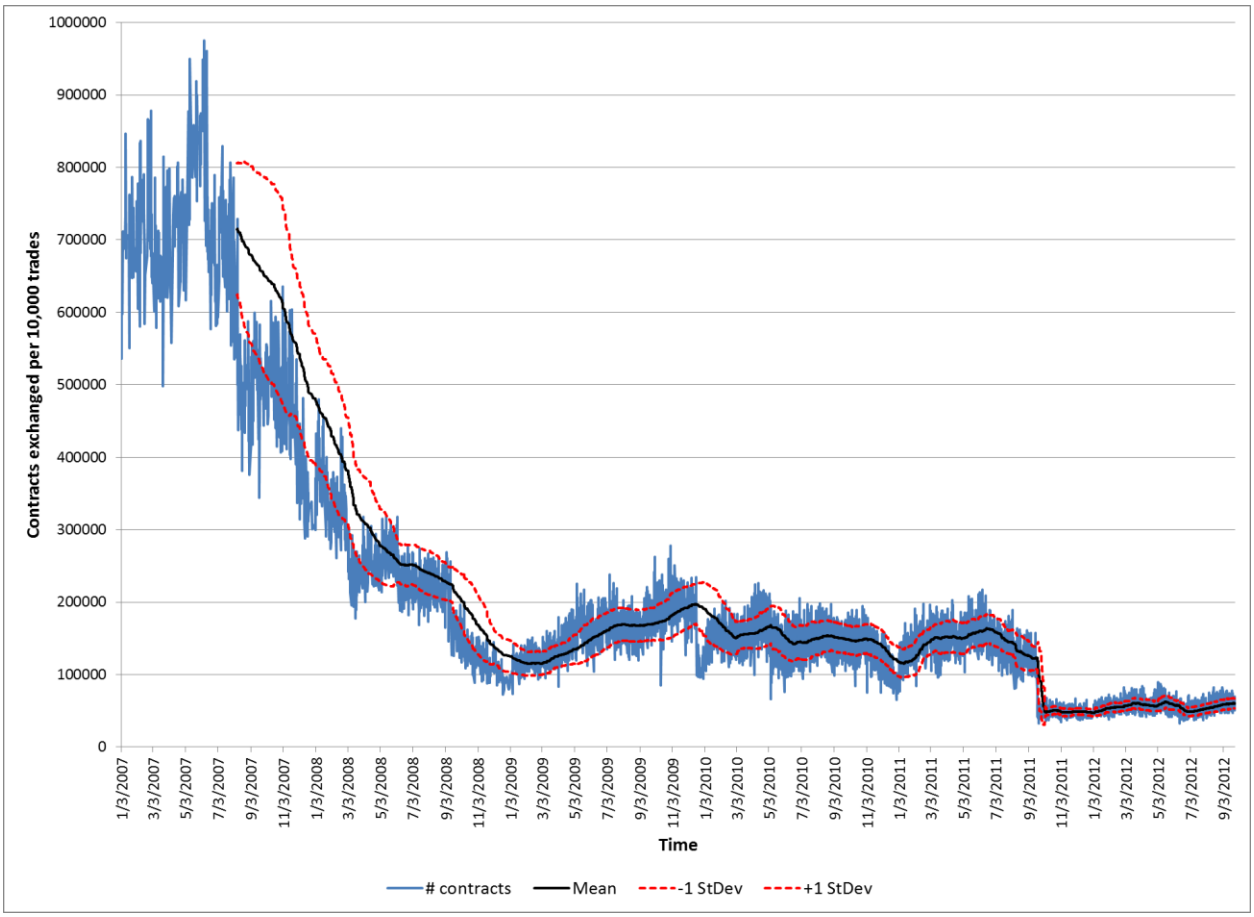
This Figure shows the average trade size in number of shares across the major equity exchanges and the Alternative Display Facility (ADF). The sample period is January 2009 – January 2014.



Source: BATS Global Markets web site

Figure 3. Trade Sizes in Treasury Bond Futures

This Figure shows the average number of contracts traded per 10,000 trades in the Treasury Bond futures contract. The sample period is from January 3, 2007 to September 3, 2012. The Treasury Bond Future trades on the Chicago Mercantile Exchange.

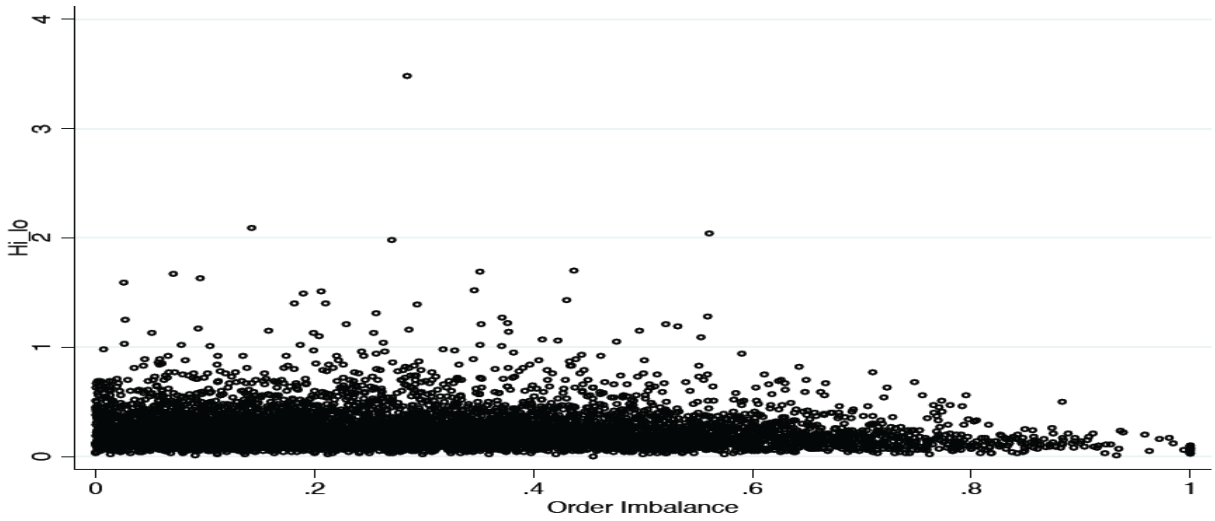


Source: CME data

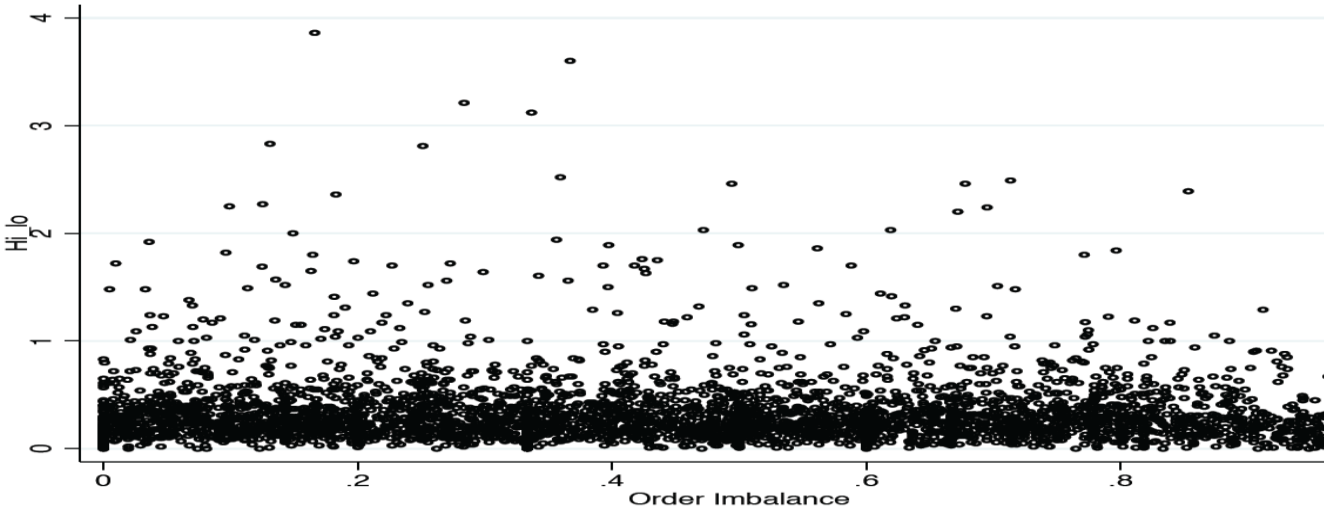
Figure 4. Trade Imbalance and Price Effects

This graph show the relationship 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 HF data set and include each trade done on the Nasdaq exchange, excluding trades done in the opening, closing, or in intraday crosses. The sample period is October 2010. Buys and sells are identified by aggressor flags in the data. Each dot presents the relationship between order imbalance and high-low price in each ten minute interval. The top panel shows trading in Apple, the middle panel shows trading in Intuitive Surgical Inc. (ISRG), and the bottom panel is trading in NX Stage Medical, Inc., (NXTM).

A. Apple (large stock)



B. Integrated Surgical, Inc. (Medium stock)



C. NX Stage Medical, Inc (small stock)