What do we know about high-frequency trading?

Charles M. Jones*

Columbia Business School

Version 3.4: March 20, 2013

ABSTRACT

This paper reviews recent theoretical and empirical research on high-frequency trading (HFT). Economic theory identifies several ways that HFT could affect liquidity. The main positive is that HFT can intermediate trades at lower cost. However, HFT speed could disadvantage other investors, and the resulting adverse selection could reduce market quality.

Over the past decade, HFT has increased sharply, and liquidity has steadily improved. But correlation is not necessarily causation. Empirically, the challenge is to measure the incremental effect of HFT beyond other changes in equity markets. The best papers for this purpose isolate market structure changes that facilitate HFT. Virtually every time a market structure change results in more HFT, liquidity and market quality have improved because liquidity suppliers are better able to adjust their quotes in response to new information.

Does HFT make markets more fragile? In the May 6, 2010 Flash Crash, for example, HFT initially stabilized prices but were eventually overwhelmed, and in liquidating their positions, HFT exacerbated the downturn. This appears to be a generic feature of equity markets: similar events have occurred in manual markets, even with affirmative market-maker obligations. Well-crafted individual stock price limits and trading halts have been introduced since. Similarly, kill switches are a sensible response to the Knight trading episode.

Many of the regulatory issues associated with HFT are the same issues that arose in more manual markets. Now regulators in the US are appropriately relying on competition to minimize abuses. Other regulation is appropriate if there are market failures. For instance, consolidated order-level audit trails are key to robust enforcement. If excessive messages impose negative externalities on others, fees are appropriate. But a message tax may act like a transaction tax, reducing share prices, increasing volatility, and worsening liquidity. Minimum order exposure times would also severely discourage liquidity provision.

*Robert W. Lear Professor of Finance and Economics, cj88@columbia.edu. I thank Larry Glosten, Michael Goldstein, Terry Hendershott, Albert Menkveld, Gideon Saar, and seminar participants at the Swedish Institute for Financial Research, the FINRA Economic Advisory Committee, and the 8th Annual Central Bank Workshop on the Microstructure of Financial Markets at the Bank of Canada for helpful discussions on some of the topics covered herein. This research was supported by a grant from Citadel LLC.

Executive Summary

U.S. equity and futures markets are highly automated, and high-frequency trading (HFT) has become a topic of regulatory focus. HFT firms typically trade hundreds or thousands of times per day for their own account, with a typical holding period measured in seconds or minutes. This paper reviews a substantial body of recent theoretical and empirical research on HFT so that researchers, practitioners, policymakers, and other interested parties can become familiar with the current state of knowledge and some of the outstanding economic issues associated with HFT. In particular, the accumulated evidence needs to be taken into account in developing equity and futures market regulations.

Based on the vast majority of the empirical work to date, HFT and automated, competing markets improve market liquidity, reduce trading costs, and make stock prices more efficient. Better liquidity lowers the cost of equity capital for firms, which is an important positive for the real economy. Minor regulatory tweaks may be in order, but those formulating policy should be especially careful not to reverse the liquidity improvements of the last twenty years.

Many HFT strategies are not new. They are simply familiar trading strategies updated for an automated environment. For example, many HFTs make markets using the same business model as traditional market-makers, but with lower costs due to automation. In fact, HFT market-makers have largely replaced human market makers. Other HFT strategies conduct cross-market arbitrage, such as ensuring that prices of the same share trading in both New York and London are the same. In the past, human traders would carry out this type of arbitrage, but the same trading strategy can now be implemented faster and at lower cost with computers.

Liquidity – the ability to trade a substantial amount of a financial asset at close to current market prices – is an important, desirable feature of financial markets. The key question is whether HFT improves liquidity and reduces transaction costs, and economic theory identifies several ways that HFT could affect liquidity. The main positive is that HFT can intermediate trades at lower cost. Those lower costs from automation can be passed on to investors in the form of narrower bid-ask spreads and smaller commissions. The potential negative is that the speed of HFT could put other market participants at a disadvantage. The resulting adverse selection could reduce market quality. There is also the potential for an unproductive arms race among HFT firms racing to be fastest.

Over the past ten years, HFT has increased sharply, and liquidity has steadily improved. But correlation is not necessarily causation. Empirically, the challenge is to measure the incremental effect of HFT on top of all the other changes in equity markets. The best papers for this purpose identify market structure changes that facilitate HFT. There have been several such changes, and the results in these papers are consistent. Every time there has been

a market structure change that results in more HFT, liquidity and overall market quality have improved. It appears that market quality improves because automated market-makers and other liquidity suppliers are better able to adjust their quotes in response to new information.

While HFT causes better market quality on average, some commentators have argued that HFT could make markets more fragile, increasing the possibility of extreme market moves and episodes of extreme illiquidity. During the May 6, 2010 Flash Crash, for example, S&P futures fell almost 10% within 15 minutes before rebounding. Some individual stocks moved far more. The CFTC and SEC were able to identify many HFT firms active during the Flash Crash, and they find that these firms initially stabilized prices but were eventually overwhelmed, and in liquidating their positions, HFT exacerbated the downturn. This appears to be a common response by intermediaries, as it also occurred in less automated times during the stock market crash of October 1987 and a similar flash crash in 1962. Thus, there does not seem to be anything unusually destabilizing about HFT, even in extreme market conditions. Short-term individual stock price limits and trading halts have been introduced since; this appears to be a well-crafted regulatory measure that should prevent a recurrence. A trading pause should give market participants a chance to re-evaluate and stabilize prices if the price moves appear unwarranted.

Regulators in the US and abroad are considering a number of other initiatives related to HFT. However, many of the issues associated with HFT are the same issues that arose in more manual markets. For example, there is concern about the effects of a two-tiered market. Today, the concern is that trading speed sorts market participants into different tiers. In the floor-based era, the concern was access to the trading floor. Many of the abuses in the floor-based era were due to a lack of competition. Now, regulators are appropriately relying on competition to minimize abuses. If there is some sort of market failure, however, then robust competition may not always be the solution, and regulation may be in order. Proposed regulatory initiatives include:

Consolidated order-level audit trails: Audit trails have always been needed for market surveillance, and robust enforcement is important to ensure investor confidence in markets. With HFT, malfeasance is possible in order submission strategies, so regulators need ready access to order-level data from multiple venues. The details turn on the costs and benefits, which are hard for an outsider to judge.

Order cancellation or excess message fees: If bandwidth and data processing requirements are overwhelming some trading venue customers, it may be appropriate for trading venues to set prices accordingly and charge the participants who are imposing those costs on others. Nasdaq is currently imposing these fees in the U.S.; it is too soon to measure the effects. However, order cancellation fees will almost certainly reduce liquidity provision away from the inside quote, reducing market depth. The current initiatives should be studied carefully before broader-based message fees are considered.

Minimum order exposure times: Under these proposals, submitted orders could not be cancelled for at least some period of time, perhaps 50 milliseconds. This would force large changes in equity markets and could severely discourage liquidity provision. The economic rationale here is particularly suspect, as the overriding goal in market design should be to encourage liquidity provision.

Securities transaction taxes: The evidence indicates that these taxes reduce share prices, increase volatility, reduce price efficiency, worsen liquidity, increase trading costs, and cause trading to move offshore.

1. Introduction

Over the past few decades, technology has transformed the trading of securities and other financial instruments. Before the advent of computers, all trading was conducted between humans, often in person on a trading floor. Back offices were filled with clerks and others to ensure that transactions were properly completed. Gradually, both the back office and the actual trading process have been transformed by automation.

Many financial markets have abandoned human intermediation via floor trading or the telephone, replacing human intermediaries with an electronic limit order book or another automated trading system. In response to an automated trading process, market participants began to develop trading algorithms. Many of these trading algorithms were designed to replicate the behavior of other humans involved in the trading process, such as agency floor brokers or proprietary market-makers. Over the past 10 years or so, these trading algorithms have been refined, computing technology continues to advance, and orders to buy and sell are appearing and matching at a faster rate than ever before.

The regulatory framework has also contributed to this automation. For example, the U.S. Securities and Exchange Commission's Regulation NMS, which was adopted in 2005, provided strong incentives for trading venues to automate, especially the New York Stock Exchange, which was the last major floor-based exchange in the U.S. Regulation NMS also encouraged competition among trading venues, competition that often took the form of technological upgrades and reductions in latency. While technology is the proximate driver, it is clear that regulation has also contributed to the current automated market structure, and regulatory policy will most certainly influence the direction of future technological developments in trading.

2. What is HFT?

According to the SEC, high-frequency traders are "professional traders acting in a proprietary capacity that engage in strategies that generate a large number of trades on daily basis." (SEC Concept Release on Equity Market Structure, 75 Fed. Reg. 3603, January 21, 2010) The SEC concept release goes on to report (p. 45) that HFT are often characterized by:

(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).

Conversations with market participants indicate that many HFT do carry substantial inventory positions overnight; otherwise, there is considerable consensus that this is a workable definition of HFT. HFT is a subset of all algorithmic trading (AT), which is generally defined as the use of a computer algorithm to make decisions about order submissions and cancellations. For example, algorithms are often used by buy-side investors to "work" large orders over time. While these algorithms are automated and often involve the rapid submission and cancellation of orders in an effort to achieve the desired execution, and AT may itself have an impact on liquidity and market quality, most of the current policy discussion focuses on proprietary traders who are trading rapidly but are not long-term shareholders.

2.A. Types of HFT Strategies

There is no single trading strategy followed by all or even most HFT. However, there are several illustrative HFT strategies, including: (1) acting as an informal or formal market-maker, (2) high-frequency relative-value trading, and (3) directional trading on news releases, order flow, or other high-frequency signals. Each category is discussed below.

2.A.1. Market-making

Market-makers simultaneously post limit orders on both sides of the electronic limit order book. They provide liquidity to market participants who want to trade immediately. Market-makers aim to buy at the bid price and sell at the ask price, thereby earning the bid-ask spread. Of course, market-makers bear the risk that they trade with, and lose money to, an informed counterparty. Thus, they have an incentive to make sure that their limit orders to buy and sell incorporate as much current information as possible as quickly as possible, so as to limit their losses to informed counterparties. As a result of this process, HFT market-makers frequently update their quotes in response to other order submissions or cancellations. HFT market-makers might also adjust quotes in response to a price move in a related ETF or futures contract. As a result of this continuous updating process, HFT market-makers tend to submit and cancel a large number of orders for each transaction.

In most U.S. equity markets, liquidity providers also earn liquidity rebates which are sometimes referred to as "maker fees". Some HFT market-makers formally register as such with trading venues. Others act as informal market-makers. This choice generally depends on the obligations and benefits associated with being a registered market-maker, and these vary across assets and across trading venues. For example, registered market-makers were exempt from the September 2008 ban on short sales in financial firms, but informal market-makers

were subject to the ban. In any case, regardless of the formal title, HFT market-makers have largely replaced traditional human market-makers, in part because they are less likely to be picked off by an informed counterparty, and in part because the use of technology results in a lower cost structure for HFT market-makers.

2.A.2. Relative value and arbitrage trading

Relative value and arbitrage trading can take many forms. A classic example is index arbitrage. S&P 500 futures are traded in Chicago on the Chicago Mercantile Exchange, while SPY is the ticker symbol for the largest exchange-traded fund (ETF) that tracks the S&P 500 index. SPY is traded on nearly every equity trading venue in the U.S. as well as several foreign trading venues. The two instruments are very similar, and their prices should move in lockstep one-for-one. If the futures price goes up due to the arrival of buy orders, but the ETF price does not move up at the same instant, HFT would quickly buy SPY, sell S&P 500 futures contracts, and lock in a small profit on the price differential between the two instruments. Naturally, these profit opportunities require rapid computer processing capability and the quickest possible link between the electronic market in Chicago and the electronic equity markets, most of which are located in New Jersey.

This example also shows the winner-take-all nature of arbitrage-oriented HFT. Continuing the above example, if one HFT arbitrageur is consistently faster than any other market participant, it will be able to quickly buy up all of the relatively mispriced shares of SPY and sell relatively mispriced S&P 500 futures contracts, thereby bringing the prices of the two instruments back into line. There will be no attractive index arbitrage trading opportunities left for a slightly slower trader.

Index arbitrage can also take place between the index products discussed above and the individual stocks that make up the index. If the S&P 500 futures price rises, for example, but there is no change in the prices of the component stocks, HFT will quickly buy shares in many or most of the underlying stocks in the correct proportions until the relative mispricing is eliminated.

Relative value trading can also take place between individual securities. The Spanish bank Banco Santander trades in Spain but has an American Depositary Receipt (ADR) that trades on the NYSE. Some companies have multiple classes of common stock, or other equity-linked securities such as convertible bonds. HFT can profit if the prices of two closely related securities temporarily diverge. Some HFT programs might trade in GM based on Ford's price moves. A quick upward move in oil futures prices might indicate a quick sale of airline stocks. A price move in a particular listed equity option might suggest a profitable trading opportunity in the underlying stock.

2.A.3. Directional trading

Some HFT firms electronically parse news releases, apply textual analysis, and trade on the inferred news. For instance, such a program might look for words like "raise" or "higher" or "increased" in close proximity to the phrase "earnings forecast," identify the company that is the subject of the news story, and in this case submit buy orders, all in milliseconds. In fact, most newswires now perform textual analyses of their own news stories prior to release. These news providers sell summary measures of the news to HFT firms, saving the firms from having to perform their own analysis and saving them precious milliseconds.

Other HFT firms trade based on order flow signals. For example, if a large buy order executes at the prevailing ask price, an HFT strategy might infer that the order submitter has substantial positive information. The HFT might then respond by buying shares itself.

A variant on this strategy is of considerable concern to large institutional traders. If a large institutional trader is gradually purchasing shares of IBM, an HFT might be able to sniff this out by identifying a sequence of large buy orders over the space of several minutes. The HFT might purchase shares of IBM, driving the price up and increasing the price that the institutional trader must pay to buy IBM shares. In fact, the HFT might eventually realize its profit by selling its purchased IBM shares to the institutional trader. To thwart these order anticipation strategies, the institutional trader may undertake great efforts to disguise its overall trading intentions. It may break up its order into very small pieces, in order to look like relatively uninformed orders from retail investors. It may trade in dark pools or use hidden orders to avoid revealing its trading intentions. Most of the time the institution will use an algorithm to do this, so we often end up with a hide-and-seek battle between computer programs.

It is important to note that this is not a new concern. In more manual markets, institutional traders were just as worried about this sort of information leakage, and they went to great lengths to disguise their large orders by working them gradually over time, perhaps by stationing a broker at the relevant trading post on the NYSE floor. Even so, institutional order execution strategies have changed substantially over the past few years as markets automate and this game of cat-and-mouse evolves. Many institutional traders express dismay over the changes they have experienced, but ultimately their trading costs are measurable, so it should be possible to assess whether these institutional traders are negatively affected by this potential aspect of HFT. In fact, as detailed later in Section 4, overall institutional trading costs have

continued to decline even as HFT becomes more prominent, suggesting that this effect, if it exists, is relatively unimportant.

2.B. Co-location

When an attractive order appears on the limit order book, only the fastest HFT can trade with this order and earn trading profits. Because of this winner-take-all characteristic of some HFT, these firms make investments in computer hardware and refine their computer algorithms in order to minimize latency, the overall time it takes to receive signals from a trading venue, make trading decisions, and transmit the resulting order messages back to the trading venue. Because light and electrical signals travel at a finite speed of 186,000 miles per second in a vacuum and somewhat slower through fiber optic cables and other media, it is important for HFT to locate its computers close to trading venue servers. For a price, electronic trading venues offer space to HFTs in their data centers, and use of these co-location services is characteristic of HFT. As discussed more later, co-location raises issues related to competition between market members, equal access to the market, and the cost of such services, but the regulatory framework that governs these services is explicitly and carefully designed to minimize these issues.

2.C. Profitability of HFT

Most HFT strategies earn only a small amount of profit per trade. Some arbitrage strategies earn profits close to 100% of the time, but many HFT strategies are based on the law of averages. Such strategies might make money on only 51% of the trades, but since these trades are conducted hundreds or thousands of times per day, they can still be consistently profitable. Hendershott and Riordan (2011) observe 25 of the largest high-frequency traders

who trade on Nasdaq during 2008 and 2009, and they find that together these HFTs earn average gross trading revenue of $2,351 per stock per day. Technology and labor costs would surely reduce these numbers considerably. These trading revenue levels are a fraction of the trading revenue earned by specialists and other market-makers 15 years ago, indicating that the compensation to trading intermediaries is much smaller than in the past.

3. The importance of stock market liquidity

3.A. Dimensions of liquidity

Liquidity is a complex concept, but it can be usefully described as the ability to trade a large amount of a financial instrument in a short amount of time at close to the current price. Thus, there are three dimensions to liquidity: price, size, and time. One source of illiquidity is explicit transaction costs, including brokerage commissions, order-handling fees, and transaction taxes. Investors who demand liquidity must buy at the ask price and sell at the bid price, so bid-ask spreads are also a component of trading costs for these traders.

Another source of illiquidity is often referred to as price pressure or price impact. If a market participants needs to sell an instrument quickly, and natural buyers are not immediately available, a market-maker may take the other side of the trade, in anticipation of laying off the position later. The market-maker is exposed to the risk of price changes while she holds the position in inventory, and thus the market-maker charges for this risk.

Price impact can also occur because a trading counterparty may have private information. For example, the buyer of a stock may worry that a potential seller has negative information about a company's upcoming earnings. Thus, the arrival of sell orders tends to drive down the share price, as other market participants infer that the sellers might be informed. In addition to private information about the fundamental value of a security, market

participants may also have private information about order flow. For example, if a trading desk knows that a hedge fund may need to liquidate a large position, the desk may sell first, lowering prices.

3.B. Measures of liquidity

Because liquidity is complex and multi-dimensional, there is no single best measure of liquidity. Researchers generally use a variety of liquidity measures. In most cases, these measures are highly correlated with each other, but inferences and conclusions can sometimes depend on the liquidity measure that is examined.

Perhaps the easiest liquidity measure to calculate and observe is the bid-ask spread. This measures the round-trip cost of a buy and sell transaction by a liquidity demander. Bid-ask spreads are usually measured as a fraction of the total amount traded (these are typically called *proportional spreads* or *relative spreads*), but sometimes bid-ask spreads are measured in cents per share.

In U.S. equity and futures markets, continuously updated bid and ask prices are electronically transmitted to investors by trading venues. These are referred to as *bid-ask quotes* or just *quotes*, and the bid-ask spread calculated using these quotes is called the *quoted spread*. Ultimately, transactions may not take place at the quoted bid or ask price. There could be a hidden order at a better price, or a market-maker may offer a better price to some incoming orders (known as *price improvement*). In both of these cases, the *effective bid-ask spread* may be narrower than the quoted spread. If the incoming order is large and exceeds the amount that is bid or offered at the quoted price, the remaining part of the order may be executed at an inferior price (sometimes known as *walking the book*). In this case, the effective bid-ask spread is actually wider than the quoted spread. Researchers generally define the

effective bid-ask spread as twice the distance between the actual transaction price and the prevailing midpoint between the quoted bid and ask prices. The multiplier is two because the distance to the actual trade price captures the cost of one side of the trade, while bid-ask spreads by convention measure a round-trip trading cost. In most markets, the effective bid-ask spread is somewhat narrower than the quoted bid-ask spreads.

While bid-ask spreads are an appropriate liquidity measure for traders who trade once in small quantities, they may not be the appropriate measure for institutions and other large traders who gradually "work" large orders over time in order to minimize their execution costs. While these traders are also concerned about bid-ask spreads, they worry at least as much about the price impact of their trades. As an institutional trader buys shares of Intel, for example, the purchases tend to drive up the Intel share price. Later purchases in the sequence tend to occur at higher prices.

These *price impacts* can be measured by looking at the response of share prices to a particular trade. However, the preferred trading cost measure for institutions is known as *implementation shortfall.* It is calculated as the average execution price for the large order compared to the price of the stock prior to the start of execution. As with spreads, implementation shortfalls are usually calculated on a proportional basis relative to the amount traded, and implementation shortfall is usually reported in basis points.

While most liquidity measures also double as measures of trading costs, there are a few liquidity measures that focus on deviations from the so-called efficient price. These measures of *price efficiency* typically measure the average size of these deviations using an econometric model that takes into account a great deal of recent order flow information (see Hasbrouck, 1992, for example). By definition, these temporary price moves are eventually reversed or eliminated, so these price inefficiencies are generally measured econometrically by looking at price reversals.

3.C. Liquidity affects stock prices

Explicit transaction costs affect share prices, because they subtract from returns every time a share of stock is bought or sold. A buyer knows that she will have to sell one day and incur transaction costs. She also knows that the investor who buys from her will have to pay transaction costs when he buys and again when he sells, and so on down the line. Thus, share prices should be reduced by the present value of all expected future transaction costs. Conversely, anything that permanently reduces transaction costs should permanently increase share prices.

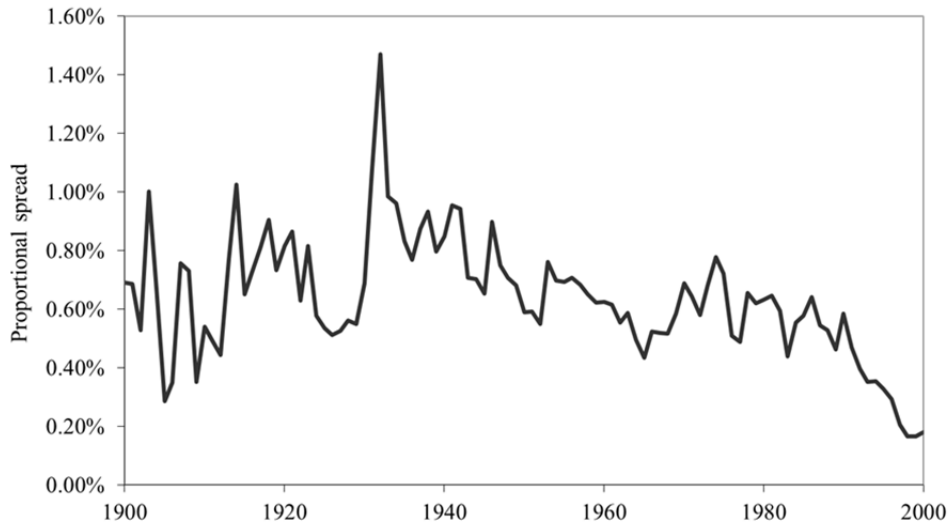
When they are justified, higher share prices are valuable for the economy, because they lower the cost of capital for firms. With a lower cost of capital, more investment projects are profitable, and firms should increase their level of investment. Greater investment should lead to higher levels of GDP and a better standard of living. This is the main channel by which HFT can have societal value. In addition, this is the main reason policymakers should work hard to facilitate low transaction costs and sustainably high levels of stock market liquidity.

Acharya and Pedersen (2005) introduce a seminal capital asset pricing model that incorporates these liquidity effects. Liquidity levels are important in their model, but so is variability in liquidity. Liquidity risk is defined as a positive correlation between illiquidity and negative stock returns, and they show that this is an undesirable risk that investors want to avoid. While not formally a part of their model, their results suggest that market structures or policies that reduce this liquidity risk would be desirable and would raise overall share prices. Thus, there is considerable justification for focusing policy on minimizing the evaporation of liquidity at exactly the wrong times.

4. Recent improvements in market liquidity

Jones (2003) finds that bid-ask spreads over the past 100 years are mostly countercyclical. Figure 1 shows the time-series evolution of quoted bid-ask spreads in Dow Jones Industrial Average stocks. The figure shows that bid-ask spreads are wide when share prices are relatively low and the economy is poor, and spreads tend to narrow when the economy and share prices improve. This pattern persists until about 1980, and then bid-ask spreads steadily narrow as part of a secular change. This probably reflects the continued adoption of technology (see for example the technological upgrades documented in Easley, Hendershott, and Ramadorai (2009)), which reduced order turnaround times and provided more timely information about conditions on the floor of the exchange, as well as modest efforts by the dominant NYSE to lessen the advantages accruing to floor participants.

Figure 1. Bid-ask spreads on Dow Jones stocks
(all DJ stocks 1900-1928, DJIA stocks 1929-present)



Chordia, Roll, and Subrahmanyam (2005), Hendershott, Jones, and Menkveld (2010), and Angel, Harris, and Spatt (2010) document the more recent narrowing of bid-ask spreads. The following figure is taken from Angel, Harris, and Spatt (2010) and shows the narrowing of bid-ask spreads over time:

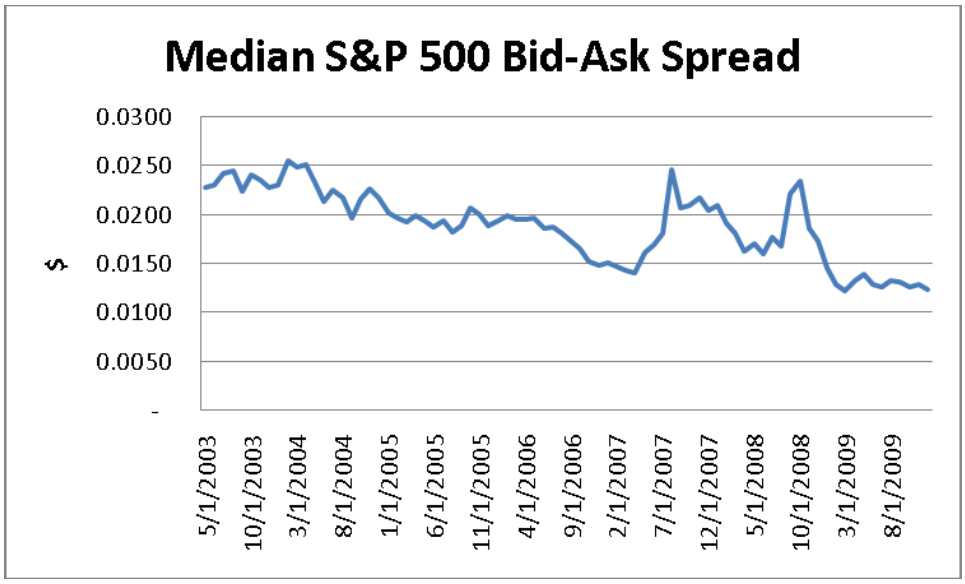


Figure 2. Median bid-ask spreads on S&P 500 stocks, 2003-2009. Source: Knight Securities via Angel, Harris, and Spatt (2010)

As discussed earlier, bid-ask spreads are the relevant measure of trading cost for investors with small orders that demand immediacy. Implementation shortfall is a better measure of trading costs for large institutional investors, and Figure 3 shows that institutional trading costs are lower in 2011 than in 2004, based on data collected by the transaction cost analysis arm of the brokerage firm ITG. Institutional trading costs are much higher during the last quarter of 2008 and the beginning of 2009, but this is almost surely due to the macroeconomic uncertainty and stock market volatility associated with the financial crisis. Trading costs drop quickly thereafter and are now back near their pre-crisis lows. While this time-series evidence cannot isolate the impact of high-frequency traders on market quality, it does indicate that US equity markets overall are doing a good job of providing liquidity to institutional traders.

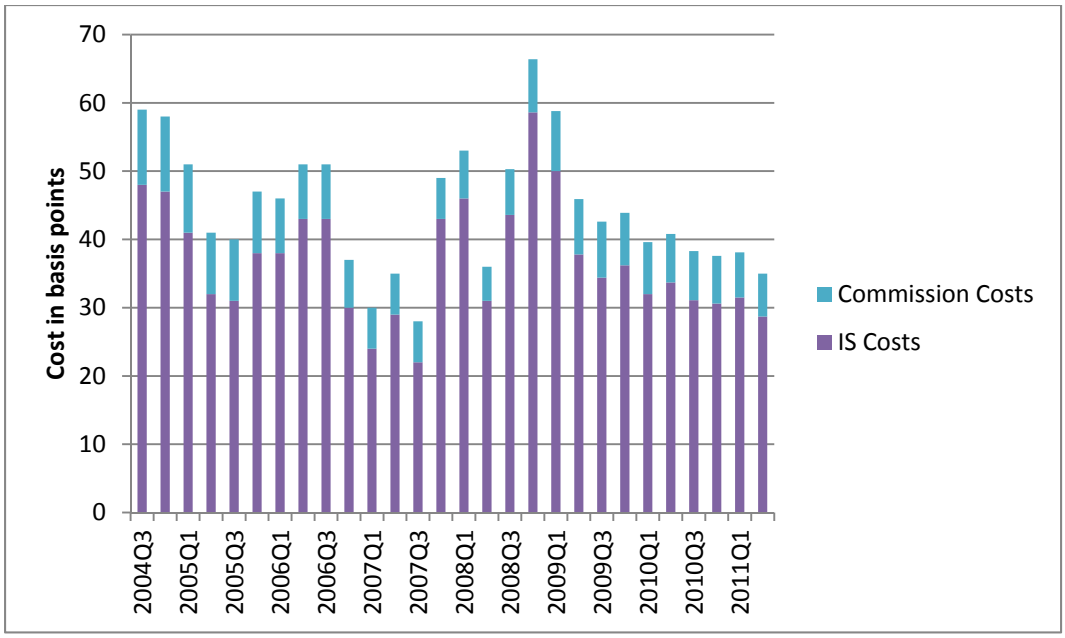


Figure 3. Average implementation shortfall in large-cap U.S. stocks. Source: ITG research reports.

Finally, French (2008) shows that commissions and other trading cost frictions have also declined over time. Figure 4 shows that overall average commissions plus market-maker revenue on equity trades in the United States have plummeted from about 1.46% of the amount traded in 1980 to 0.11% in 2006, the last year for which data are available.

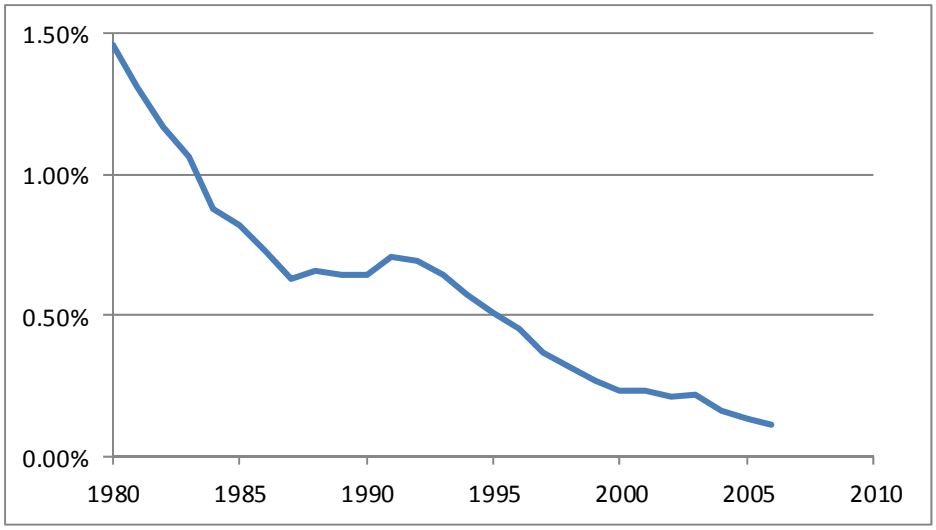


Figure 4. Average commissions plus markups on U.S. equity trading, expressed as a percentage of volume transacted. Source: French (2008).

There are many reasons that liquidity has improved and trading costs have declined over time. Commissions were deregulated in the early 1970's and quickly began to plummet. Stock exchanges gradually adopted new technology and reduced costs, passing those lower costs on to broker-dealers and investors in the form of lower trading fees. The minimum price increment was $0.125 per share before 1997, which meant that the minimum bid-ask spread was also $0.125. The minimum tick is now $0.01. In the mid 1970's, Congress enacted the National Market System, which was explicitly designed to encourage competition between various trading venues. Eventually, competing trading venues were able to draw trading volume away from the incumbent NYSE and Nasdaq systems. Regulation NMS, which came into effect in 2006, was an important part of encouraging this competition.

As exchanges turned to automation, it became possible to replace humans with machines at various points in the trading process. Floor brokers, who would stand at the NYSE trading post and work customer orders over the course of a trading day, were gradually replaced by algorithms that were responsible for the gradual execution of large customer orders over time. Human market-makers were gradually replaced by automated market-makers. And computer programs were quicker to spot arbitrage opportunities and could quickly do the required calculations to identify mispriced securities. In these cases, computers were cheaper than humans, and the resulting cost efficiencies would be expected to be passed on to investors.

Ultimately, it is somewhat difficult to pin down the contribution of each of these changes, but researchers have studied each of their effects. This paper focuses on the potential contribution of high-frequency trading in all of its various forms to the recent changes in liquidity and trading costs.

5. Economic issues surrounding HFT

Basic textbook economics can be easily applied to HFT. For example, if there is competition in the provision of liquidity, the more efficient provider of liquidity is likely to win out. This benefits consumers, who in this case are investors with a desire to trade.

Liquidity is also affected by asymmetric information across market participants. If better-informed traders interact with less-informed traders, the better-informed traders are likely to buy low and sell high, earning profits, while the less-informed traders are likely to buy high and sell low, generating losses. This is often referred to as adverse selection, because the less-informed traders end up with relatively unattractive trades. In general, holding everything else equal, if HFT increases adverse selection, liquidity worsens.

If HFT are indeed better informed due to their speed in processing information, there is also an upside, because their trading contributes to price discovery. Stock prices are more efficient because they reflect more information more quickly, and this can be valuable to all investors. For example, as discussed earlier index arbitrage activity by HFT ensures that the price of a basket of stocks reflects the prices of the underlying stocks. An investor who wants to hold a broad market index such as the S&P 500 can purchase SPY (the biggest ETF that tracks the S&P 500) secure in the knowledge that the price of the ETF closely reflects the trading price of the underlying stocks. This also makes SPY an effective hedging tool for those trying to limit their stock market risk. SPY is by far the most actively traded security on US stock exchanges, and its volume has increased dramatically over the past several years, suggesting that the price discovery due to HFT and other arbitrageurs is quite valuable to investors. The resulting trading volume and high levels of liquidity make SPY attractive to

more and more investors, and might encourage them to take bigger positions or otherwise make greater use of the ETF.

Asymmetric information and price discovery have long been of interest in economic models of trading, and some older economic models are explicitly designed to consider these types of tradeoffs. These models can be easily applied to HFT, because the basic economics of market-making and the effects on markets of differentially informed investors are the same whether the market is manual or automated.

Glosten and Milgrom (1985), for example, is one of the most important theoretical market microstructure models, and one of the first to consider adverse selection in trading. In their model, competitive market-makers transact with potentially informed traders. They have in mind a market where investors transact with dealers by phone, but the results are not dependent on those particular details. They show that the equilibrium bid-ask spread depends on the cost of processing trades and the amount of adverse selection, specifically the information advantage that the informed traders have over market-makers. The model can be applied directly to the current automated trading structure. For example, if high-frequency market-makers have lower costs than traditional market-makers, Glosten-Milgrom implies that this cost reduction should manifest itself in narrower bid-ask spreads. In addition, if high-frequency market-makers are able to incorporate more information in setting a quote, this should reduce their information disadvantage, and this too should result in a narrower bid-ask spread. Of course, not all HFT are market-makers. If HFT make the adverse selection problem worse for any reason, spreads widen in the Glosten-Milgrom model.

Hirshleifer (1971) shows that there are strong incentives to collect private information and trade on that information, and the private value of that information can far exceed the social value of that information. This suggests the possibility that HFT might overinvest in

technology, infrastructure, and information production relative to social optimum. It is important to note that the Hirshleifer model is explicitly constructed so that there are no benefits to having more efficient or accurate prices. If price discovery is valuable, HFT investments in technology can be socially desirable, reflecting healthy competition rather than an unproductive arms race.

With the recent interest in HFT, there have been a number of recent theoretical models focused explicitly on HFT. Typically in these models, buyers and sellers who arrive at the market at different times are unable to trade with each other, and by holding inventory for a short period of time, HFT enables gains from trade. In most of these models, the downside of HFT is that their speed, or the information they collect and use for trading, increases adverse selection, thereby worsening liquidity.

In the model of Biais, Foucault, and Moinas (2011), the intermediation function provided by HFT helps traders find counterparties, leading to gains from trade. However, HFT can trade on new information more quickly, generating adverse selection costs. In addition, HFT requires significant fixed investments in technology. In their model, this means that only sufficiently large institutions are likely to make these fixed investments. Smaller firms and investors are left to bear the adverse selection costs from HFT. Finally, they model the arms race feature of HFT. They find multiple equilibria in their model, some of which exhibit socially inefficient overinvestment in HFT.

There is a similar trade-off in Jovanovic and Menkveld (2011). In their model, HFT can update limit orders quickly based on new information. As a result, HFT can avoid some adverse selection, and HFT can provide some of that benefit to uninformed investors who need to trade. Some of these trades might not have occurred otherwise, in which case HFT can improve welfare. However, if the natural buyers and sellers do not have much of an adverse

selection problem, the HFT can introduce one, reducing welfare compared to a world where buyers and sellers meet directly on a limit order book.

In Martinez and Rosu (2011) and Foucault, Hombert, and Rosu (2012), the focus is on HFTs that demand liquidity. These HFTs receive a stream of signals about changes in the value of an asset. As a result, HFTs generate a large fraction of the trading volume and price volatility. In Martinez and Rosu (2011), this volume and volatility is desirable, as HFT makes market prices extremely efficient by incorporating information as soon as it becomes available. Markets are not destabilized, as long as there is a population of market-makers standing ready to provide liquidity at competitive prices. In Foucault, Hombert, and Rosu (2012), an HFT obtains and trades on information an instant before it is available to others. This imposes adverse selection on market-makers, so liquidity is worse, and prices are no more efficient.

Other related theoretical models include Pagnotta and Philippon (2011), who focus on the investment in speed made by exchanges in order to attract trading volume from speed-sensitive investors. Cartea and Penalva (2011) simply assume that HFT constitute an unnecessary extra layer of intermediation between buyers and sellers. Not surprisingly, they find that HFT reduces welfare. Moallemi and Saglam (2012) argue that a reduction in latency allows limit order submitters to update their orders more quickly, thereby reducing the value of the trading option that a limit order grants to a liquidity demander.

As noted earlier, the most common theme in these models is that HFT may increase adverse selection, which is bad for liquidity. The ability to intermediate traders who arrive in the market at different times is usually good for liquidity. Unfortunately, none of these models addresses the fact that HFT market-makers are simply more efficient liquidity providers because they have replaced humans with technology. In some sense, this cost reduction is too

simple to add to a model, but its omission is particularly unfortunate, because these cost reductions are likely to be an important source of HFT's liquidity benefits.

6. Empirical Studies Related to HFT

In the past few years, there have been a number of studies of HFT and algorithmic trading more generally. Some researchers have been able to identify a specific HFT in the data. Other researchers are able to identify whether a trade is from HFT or AT. Still others have been able to draw inferences about changes in HFT or AT based on changes in order submission behavior over time. With data on specific or aggregated HFT, it is possible to describe patterns in HFT order submission, order cancellation, and trading behavior. It is also possible to see whether HFT activity is correlated with bid-ask spreads, temporary and/or permanent volatility, trading volume, and other market activity and market quality measures. But correlation is not causation. It could be that HFT is simply responding to changes in market conditions, not causing them. We do not get to observe the counterfactual, which in this case would be a trading world without HFT, but the most instructive papers at this point are those that do the next best thing: study the effects of a specific change in market structure that either helps or hinders HFT. These papers all come to the same conclusion: HFT and AT improve market quality.

6.A. Papers that study specific market structure changes

Hendershott, Jones, and Menkveld (2011) study the implementation of an automated quote at the New York Stock Exchange in 2003. Before "autoquote," specialist clerks manually updated the best bid and offer prices at the NYSE. The automated quote provided more timely insight into market conditions on the floor, and it enabled algorithms and HFT to submit and

cancel orders and to quickly see those orders reflected in the NYSE's disseminated quote. HJM measure the amount of algorithmic trading as the amount of electronic message traffic (which consists of order submissions, order cancellations, and trade executions) per unit of trading. Autoquote was rolled out gradually on the NYSE, starting with a few stocks and adding additional stocks in several subsequent waves. This allows HJM to use the introduction of autoquote as an instrumental variable for the amount of algorithmic trading in a differences-in-differences approach that compares the market quality of "treated" autoquoted stocks to "untreated" stocks still using manual quote dissemination. HJM find that the implementation of autoquote is associated with an increase in electronic message traffic and an improvement in market quality. After the implementation, effective spreads narrow, adverse selection is reduced, and more price discovery takes place via quotes vs. trades. The effects are concentrated in large-cap firms; there is little effect in small-cap stocks, though their approach has relatively little statistical power because most small-cap stocks adopted autoquote in the last wave.

Overall, HJM convincingly show that, at least for the NYSE-centric U.S. market structure of 2003 and at least for large-cap stocks, an increase in algorithmic trading causes an improvement in stock market quality.

Jovanovic and Menkveld (2011) and Menkveld (2012) study the July 2007 entry of a high-frequency market-maker into the trading of Dutch stocks. The main market for these stocks was Euronext; the new market-maker set up shop on the competing Chi-X market. Chi-X is distinguished by fast execution and a fee structure that pays rebates to liquidity providers, which makes it an appealing venue for a high-frequency market-maker. Due to Dutch trade reporting requirements, all of the trades of this new market-maker can be observed, so this paper is important in providing insights on how high-frequency market-makers trade. The

new market-maker trades on both Chi-X and Euronext, but is fairly dominant on Chi-X, participating in 49.9% of all trades there. Exactly 80.0% of the new market-maker's Chi-X trades are the result of passive orders. For the 12 stocks studied over a 71-day interval, the new market-maker has exactly zero change in inventory on 47% of days, when positions are aggregated across Euronext and Chi-X. In fact, this market-maker's inventory position mean reverts rapidly during the day, typically crossing zero tens of times during a trading day. Menkveld (2012) argues that competition between trading venues facilitated the arrival of this high-frequency market-maker and HFT more generally, and he shows that on average this market-maker earned €1.55 per trade on the spread but lost €0.68 on its resulting inventory position, a pattern of profit and loss that is quite typical of a classical market-maker.

Jovanovic and Menkveld also examine the effect of entry on liquidity and market quality. The basic empirical design in the paper is differences-in-differences. Bid-ask spreads and other market quality measures on the "treated" Dutch stocks are compared to market quality measures for a set of Belgian control stocks. The Belgian stocks were not available for trading on Chi-X at the time and thus were unaffected by the entry of the high-frequency market-maker. Compared to the "untreated" Belgian stocks, effective bid-ask spreads on the Dutch stocks are about 15% narrower. High-frequency market-maker entry is also associated with 23% less adverse selection, perhaps because this market-maker updates its quotes more quickly in response to new information arriving in the market. Volatility is measured using 20-minute realized volatility, and it is unaffected by the entry of the high-frequency market-maker.

Overall, this study is important because it gives us a direct view of how a high-frequency market-maker trades, and it clearly shows that the addition of a high-frequency market-maker improves market quality in this context. This is a classic case where more

competition leads to lower prices. Here, more competition in market-making leads to narrower bid-ask spreads and reduced trading costs for other investors.

Riordan and Storkenmaier (2012) examine the effect of a technological upgrade on the market quality of 98 actively traded German stocks. In April 2007, the Deutsche Boerse made a series of system upgrades with the sole purpose of reducing the latency of its electronic limit order book. As a result, the time between order entry and confirmation was reduced from about 50 milliseconds to about 10 milliseconds. Not surprisingly, lower latency increases the rate of order submission substantially, from an average of 2.81 quote updates per 10,000 euros of trading volume to an average of 4.56 quote updates post-upgrade. Effective spreads narrow, going from an average of 7.72 basis points pre-upgrade to 7.04 basis points afterward. Average price impacts drop considerably with the upgrade, from 6.87 basis points to 2.65 basis points, and realized spreads increase from an average of 0.97 basis points to 4.45 basis points. This suggests that the ability to update quotes faster helps liquidity providers minimize their losses to liquidity demanders. It also appears that more price discovery is taking place via quote updates vs. trades. These results are consistent with results in Hendershott, Jones, and Menkveld (2011). Both sets of results are consistent with some of the winner-take-all aspects of market-making, because the results suggest that the fastest high-frequency market makers are able to earn greater trading revenues when the market's latency is reduced.

Boehmer, Fong, and Wu (2012) examine international evidence on electronic message traffic and market quality across 39 stock exchanges over the 2001-2009 period. The most interesting empirical part of the paper investigates the effect on market quality of offering co-location facilities to algorithmic and high-frequency traders. They find that co-location increases AT and HFT, and the introduction of co-location improves liquidity and the informational efficiency of prices. They claim that co-location increases volatility, but a more

accurate characterization of their results is that when co-location is introduced, volatility does not decline as much as would be expected based on the observed narrower bid-ask spreads.

Finally, Gai, Yao, and Ye (2012) study the effect of two recent 2010 Nasdaq technology upgrades that reduce the minimum time between messages from 950 nanoseconds to 200 nanoseconds. These technological changes lead to substantial increases in the number of cancelled orders without much change in overall trading volume. There is also little change in bid-ask spreads and depths. This suggests that there may be diminishing liquidity benefits to faster exchanges.

The other market structure changes that have been closely studied all yield the same conclusions. Increased automation at the NYSE enhanced market quality in large stocks, the entry of an HFT market-maker on Chi-X narrowed spreads, as did a reduction in latency on the Deutsche Boerse. When co-location is offered at different times across many different stock markets, liquidity and the informational efficiency of prices both improve. There is also a clear mechanism for the liquidity improvements. After these changes, the evidence indicates that more price discovery takes place via quotes rather than trades, suggesting that market quality improves because automated liquidity suppliers are better able to adjust their quotes.

6.B. Studies that can distinguish between humans and machines

There are also a number of papers that use data where it is possible to distinguish between different types of traders. For example, Hendershott and Riordan (2012) use exchange classifications distinguish AT from orders managed by humans.[1] Brogaard (2011a, 2011b, 2012) and Hendershott and Riordan (2011) work with Nasdaq data that flags whether trades involve HFT. Baron, Brogaard, and Kirilenko (2012) use account-level trade-by-trade data on

[1] While this paper focuses on equity markets, Chaboud, Hjalmarsson, Vega, and Chiquoine (2009) characterize AT in the foreign exchange market. Their results are similar.

the e-mini S&P 500 futures contract, and they classify traders into various categories, including passive HFT and aggressive HFT. Benos and Sagade (2012) conduct a similar analysis using UK equity data. These different datasets yield considerable insight into overall HFT trading behavior. However, these papers are less well-suited to identify the causal effects of HFT on market quality.

During 2008, the Deutsche Boerse offered fee rebates to algorithmic traders. In order to administer these fee rebates, the exchange internally identified the automated traders. Hendershott and Riordan (2012) have data from the Deutsche Boerse distinguishing AT from human order submitters. They find that AT concentrates in smaller trade sizes, while large block trades of 5,000 shares or more are predominantly originated by humans. AT more actively monitor market liquidity than human traders. AT consume liquidity when bid-ask spreads are relatively narrow, and they supply liquidity when bid-ask spreads are relatively wide. This suggests that algorithmic traders may reduce the variability in market quality, providing a more consistent level of liquidity through time.

Hendershott and Riordan (2011) and Brogaard (2011a, 2011b, 2012) use 2008-2009 millisecond-stamped data on all Nasdaq trades in 120 stocks that flags whether each of the parties to the trade is an HFT. Brogaard also has similar data for trades on BATS. The firm behind each trade is categorized as HFT if it engages only in proprietary trading, its net position often crosses zero during the day, and its non-marketable orders are typically short-lived. This excludes most large broker-dealers such as Goldman Sachs and Merrill Lynch, who may undertake some high-frequency proprietary trading but also act as brokers for customers. In addition, some HFT is routed through these larger broker-dealers, and as a result these trades would not be classified as HFT in the dataset. Ultimately, the HFT designation is

applied to trades made by 26 different independent proprietary trading firms, and thus the trades flagged as HFT in the Nasdaq dataset should be considered a subset of all HFT.

Hendershott and Riordan (2011) find that HFT accounts for about 42% of (double-counted) Nasdaq volume in large-cap stocks but only about 17% of volume in small-cap stocks. Brogaard (2011a) similarly finds that 68% of trades have an HFT on at least one side of the transaction, and he also finds that HFT participation rates are higher for stocks with high share prices, large market caps, narrow bid-ask spreads, or low stock-specific volatility. Recall that these are lower bounds on the prevalence of HFT, because Nasdaq is only able to flag a subset of overall HFT activity. Brogaard also finds that there is a bit less HFT at the very beginning and end of the trading day. In large stocks, HFT demand and supply liquidity on Nasdaq about equally. In small-cap stocks, HFT tends to demand liquidity more often. In these stocks, HFT accounts for 23% of liquidity demand but only 10% of liquidity supply. Brogaard finds that HFT liquidity suppliers are more likely to be at the inside in large-cap stocks.

Hendershott and Riordan (2011) estimate a state-space model that decomposes price changes into permanent and temporary components, and measures the contribution of HFT and non-HFT liquidity supply and liquidity demand to each of these price change components. For the permanent component of prices, we expect net buying by HFT liquidity demanders to be positively correlated with future price changes, reflecting their information content. Similarly, we would expect net buying by HFT liquidity suppliers to be negatively correlated with future price changes, reflecting adverse selection from better-informed liquidity demanders. This is borne out by the data. More interesting is the relationship between HFT and the temporary component of prices. It turns out that when HFT initiate trades, they trade in the opposite direction to the transitory component of prices. That is, when prices deviate from fundamental values, HFT initiate trades to push prices back to their efficient levels. These

prices return to efficient levels within about 30 seconds. Thus, HFT contributes to price discovery and contributes to efficient stock prices. The results are very similar when days are separated into higher volatility and lower volatility days.

Brogaard (2011b) estimates a vector autoregressive permanent price impact model and finds that HFT liquidity suppliers face less adverse selection than non-HFT liquidity suppliers, suggesting that they are somewhat judicious in supplying liquidity. Brogaard (2011b) also finds that prices do not exhibit any short-run overreaction following an HFT trade that demands liquidity. This is consistent with the Hendershott and Riordan results and implies that HFT makes prices more efficient.

Furthermore, there is no evidence in either Brogaard (2011b) or in Brogaard (2012) that HFT increases market volatility. Brogaard (2012) examines the 2008 temporary ban on short sales in financial stocks that was in place for about three weeks in September and October. While registered market makers were exempt from the ban (see Boehmer, Jones, and Zhang (2012) for more details), most HFT at the time was conducted by entities that were not registered as market-makers, and thus the shorting ban sharply limited HFT trading opportunities. Brogaard uses a triple-difference approach and finds that financial stocks with the biggest decline in HFT as a result of the ban experienced the biggest increases in volatility, suggesting that HFT is important in limiting excess volatility.

Finally, it is possible to use these data to measure overall profitability of these HFT. Hendershott and Riordan (2011) find that trading revenue per day for these HFT firms (net of all rebates and take fees) averages $2,351 per stock, $6,643 per large-cap stock. Dividing by the amount of observed HFT activity, this amounts to about 4 cents per $1,000 traded by an HFT, or 0.4 basis points. Brogaard (2011a) extrapolates from his data to estimate total HFT revenue of about $3 billion per year, which sounds large, but he then points out that HFT

trading revenues are an order of magnitude smaller than the trading revenues earned by NYSE specialists in 2000, before the advent of most HFT. Baron, Brogaard and Kirilenko (2012) find that HFT in the e-mini S&P 500 futures contract earn average trading revenue of $1.11 per contract, which is about 0.002% of the value of the stocks underlying each contract and is far smaller than the minimum price increment or minimum tick of $12.50 per contract.

Overall, Hendershott and Riordan (2011) find that HFT has a beneficial role in the price discovery process. However, the information possessed by HFT is short-lived, typically less than 30 seconds. As Hendershott and Riordan write, "if this information would become public without HFT, then the potential welfare gains may be small." Still, there is no evidence that HFT contribute to market instability in prices. In fact, HFT reduce transitory pricing errors, thereby stabilizing prices, and they do this on low-volatility and high-volatility days during a relatively turbulent period.

6.C. Other related work

Other papers cannot observe which trades and orders are due to AT or HFT and must try to identify these trades or orders indirectly. Notable among these papers is Hasbrouck and Saar (2011), who examine Nasdaq order-level data (also known as ITCH data) from 2007 and 2008 and characterize some of the order submission and cancellation patterns in the data. They find evidence that some market participants (presumably HFT) are able to respond to an incoming order in about 2 milliseconds. They also examine sequences of cancel-and-replace behavior: the cancellation of a resting limit order followed almost immediately by submission of a new order for the same number of shares but at a new price. They conjecture that these sequences originate from high-frequency traders and other algorithmic traders, and in fact they coin the phrase "low-latency trading" to describe trading behavior that is characterized by rapid

responses (on the order of milliseconds) to incoming order flow and other available information.

Hasbrouck and Saar then go on to investigate the relationship between low-latency trading and market quality. They use an instrumental variable approach to determine whether more low-latency trading causes market quality to improve. The instrument is the number of cancel-and-replace sequences in other stocks at a given time, indicating more overall low-latency activity. They find that when there are more of these cancel-and-replace sequences, market quality is better. The instrument may not be exogenous, and there is some concern about a reverse causality story, but there is nothing in the empirical data to suggest that low-latency activity worsens market quality.

Zhang (2010) essentially defines HFT as all trading activity that is not captured in quarterly institutional holdings data obtained from Section 13(f) SEC filings. This is a much broader definition of HFT than used in other studies, as his measure is designed to capture trades with holding periods shorter than a calendar quarter. Even this cannot be accomplished perfectly, because small institutions are not required to file a 13(f), the reports do not reflect short positions, and so on. He investigates volatility and price overreactions using a triple-difference approach. This methodology essentially measures volatility during the 1995-2009 period (his HFT interval) compared to volatility during the non-HFT period of 1985-1994, and after controlling for many other stock and firm characteristics, he finds that stocks with more HFT by his measure experience a larger increase in volatility over this timeframe. Stocks with more HFT also appear to overreact to quarterly earnings news. Unfortunately, the measure of HFT is so poor, and there are so many confounding changes in market conditions over the time period of the study, that it is virtually impossible to draw any conclusions about HFT from this study.

Egginton, Van Ness, and Van Ness (2012) examine short episodes of intense quoting activity in 2010. They identify 1- to 10-minute periods with intense quoting activity that is more than 20 standard deviations above normal for a particular stock. These episodes are fairly common; they occur several hundred times per day in a wide variety of stocks. They find that these periods are associated with wider bid-ask spreads and greater price volatility. However, they are unable to determine whether algorithms and HFT are causing liquidity to worsen, or whether the illiquidity simply reflects the presence of private information during these episodes, with heightened quoting activity a natural response to the information environment.

Overall, the evidence strongly indicates that HFT is good for average market quality, with more discernible positive effects in large-cap stocks. The next section examines whether HFT contributes to market quality during unusually volatile times. Market quality in stressed conditions may be of particular concern to investors and regulators, because in those conditions some market participants may have acute trading needs, and illiquidity may be particularly undesirable. Thus, it may be important to consider potential links between HFT and the robustness of market quality.

7. HFT, the flash crash, and tail behavior

On May 6, 2010, U.S. stock markets were indeed stressed. Unfortunately, market quality was far from robust that day. Liquidity temporarily disappeared and share prices fluctuated wildly as U.S. stocks experienced what has become known as the "flash crash." As the trading day unfolded, stocks gradually fell on Euro-zone macroeconomic concerns, and by 2:30 PM ET, the Dow Jones Industrial Average was down 2.5%. Suddenly, stocks began to fall much more rapidly. During the 13 minute period from 2:32 PM to 2:45:27 PM, the front-month June 2010 E-mini S&P 500 futures contract fell 5.1%. In the next *second*, a flurry of sell

orders caused the futures price to drop an additional 1.3%. At this point, buy orders on the CME's electronic limit order book were so sparse that the next marketable sell order would have caused the futures contract to drop by an additional 6.5 index points, from 1056.00 to 1049.50. This large potential price move caused the CME trading software to impose a 5-second trading halt. Within seconds of resuming trading, the E-mini ascended rapidly, rising 6.4% by 3:06 PM ET and virtually erasing the afternoon's sharp drop. Figure 5 shows the behavior of stock index futures on that day.

Figure 5. E-mini S&P 500 futures prices during the flash crash of May 6, 2010.

Some individual stocks experienced even bigger gyrations. Procter and Gamble fell by more than one-third, and consulting firm Accenture momentarily traded at a share price of $0.01. Some stocks even traded at astronomically high prices during the rebound period. For

example, Apple Computer traded at a share price of $100,000 at 3:29:30 PM. A total of 20,761 trades at the most extreme prices were later cancelled, but trades that were less than 60% away from a stock's 2:40pm share price were allowed to stand. Transactions occurred at these extreme prices because all other liquidity supply disappeared during the flash crash, leaving only market-maker "stub quotes" with placeholder bids and offers at extreme prices. Since a great deal of liquidity is supplied by HFT in current markets, it is tempting to blame HFT for the apparent lack of liquidity supply during the flash crash.

Domowitz (2010), a former academic now at ITG, points out that a similar Flash Crash occurred more than 50 years ago, on the afternoon of May 28, 1962. The Dow Jones Industrial Average fell sharply in 20 minutes that day, with some stocks plummeting more than 9 percent in less than 12 minutes. That temporary dislocation occurred in a centralized market system with manual trading on the floor of the New York Stock Exchange. The SEC produced an investigative report that highlighted the role played by NYSE specialists. As the decline unfolded, specialists either stepped aside or actively sold shares. Electronic market-makers also retreated on May 6, 2010, but Domowitz writes that "this has more to do with a human reaction to step aside when markets accelerate sharply than it does with market structure."

The joint CFTC/SEC report on the flash crash states that its proximate cause was a 2:32PM order from a mutual fund complex to sell a total of 75,000 e-mini S&P 500 futures contracts as a hedge to an existing equity position. Orders this large appear in CFTC data on e-mini S&P 500 futures contracts only a couple of times per year. An algorithm began to execute this large order in the market, with the algorithm parameters set so that the mutual fund seller would be about 9% of the trading volume. The algorithm continued to submit a large number of sell orders even as the S&P futures price began to drop sharply. The order continued to execute during the price rebound, and it ultimately finished executing in about 20

minutes. Months earlier, a similar order from the same seller took more than five hours to execute, highlighting the rapid pace of the May 6 selling program. On the afternoon of May 6, given the recent stock market declines and the general macroeconomic backdrop, potential buyers observing the pace of orders to sell would naturally worry that they were facing one or more well-informed sellers with extremely negative information about fundamentals.

Around this time, there were also market infrastructure problems at NYSE Arca. According to the joint CFTC/SEC report (p. 77), the NYSE set quote traffic records during the flash crash, and this caused significant delays in the dissemination of trades and quotes for about half of the stocks traded there. In particular, there were delays averaging as much as 20 seconds in disseminating quotes to the Consolidated Quote System (CQS). There were also delays in the NYSE's proprietary data feeds. This may have caused uncertainty among other market participants about systems integrity, and it may have even caused some liquidity suppliers at other markets to step back.

While HFT may have contributed a good part of the electronic message traffic that overwhelmed the NYSE that afternoon, it is also important to investigate the trading behavior of HFT on that day, and Kirilenko, Kyle, Samadi, and Tuzun (2011) study HFT in the e-mini S&P 500 futures market during the flash crash. They have audit trail data for all 15,000 accounts that traded the e-mini that day, and they partition these accounts into six categories, including a category for HFT. KKST find that HFT did not trigger the Flash Crash, but their responses to the unusually large selling pressure on that day exacerbated the decline and worsened market volatility. In particular, as a large number of aggressive sell orders arrived after 2:30PM, HFT initially provided liquidity. Within a few minutes, possibly because they were overwhelmed by selling pressure, HFT's reversed course and aggressively liquidated their long positions, and thereby contributing to the price decline.

The joint CFTC/SEC report also examines the aggregate equity market trading behavior of 17 HFT firms during the flash crash. These 17 HFT firms accounted for around half of equity trading volume during the afternoon. In aggregate, these firms sold rapidly in the 15-minute interval between 2:30pm and 2:45pm, with net selling of $1.158 billion during that interval. As the report notes, some of this aggressive selling could be due to cross-market strategies involving the purchase of futures and the sale of equities. However, the CFTC/SEC report concludes that "it appears that the 17 HFT firms traded with the price trend on May 6 and, on both an absolute and net basis, removed significant buy liquidity from the public quoting markets during the downturn."

Easley, Lopez de Prado, and O'Hara (2011, 2012) do not have access to account-level data for the flash crash, but they use intraday data to calculate an order imbalance measure that they call the volume-synchronized probability of informed trading (VPIN). They find that order imbalances were especially severe in the minutes just prior to the flash crash, supporting the KKST conclusion that HFT and other trading intermediaries were overwhelmed by selling pressure.

Other papers related to the flash crash include Madhavan (2012), who finds that securities with more fragmented trading and quoting behavior suffered more extreme price moves during the 2010 Flash Crash. He argues that the combination of fragmentation and high-frequency traders can cause the withdrawal of liquidity in times of market stress. McInish, Upson, and Wood (2012) focus on the use of intermarket sweep orders (ISOs) to demand liquidity during the flash crash. These orders are often used to simultaneously access liquidity at multiple trading venues, and they find a sharp increase in the use of ISOs during the flash crash, consistent with some of the aggressive selling documented earlier.

While some observers suggested greater obligations for market-makers, experience in other rapid downdrafts, including the stock market crash of October 1987, when Nasdaq market-makers and others refused to answer their phones or provide market-making activity, indicates that market-makers will almost always choose to withdraw from the market in the face of such extreme volatility. Regulators and market participants also noticed that a 5-second pause in e-mini S&P futures trading was sufficient to end the flash crash slide, and their attention quickly turned to short-term trading halts.

Single-stock circuit breakers were phased in beginning about a month after the flash crash (see SEC Releases 34-62251 and 34-62252, both dated June 10, 2010). These force a 5-minute trading halt if the transaction price of an individual stock moves by more than 10% within a 5-minute period. The trading pause is designed to give market participants time to consider available information more fully and provide stabilizing liquidity if the large price move does not appear warranted based on fundamental information. From an economic point of view, the pause is designed to limit the possibility of extreme adverse selection for market-makers, give liquidity providers a chance to collect more information, and then resume trading via an auction where buyers and sellers can meet directly with less need for an intermediary. At that point, liquidity providers should be able to resume normal market-making activity.

The SEC, the national exchanges, and FINRA should be commended for quickly agreeing to and adopting single-stock circuit-breakers. While there have been no formal studies, the circuit-breakers seem to have assuaged investor fears about the wholesale disappearance of liquidity over a short period of time. Though most observers believe that these single-stock circuit breakers have generally worked well, they are sometimes triggered by a single erroneous trade on one trading venue, at a time when the market in that stock was operating in an orderly fashion on all other venues. Thus, the SEC, the national exchanges,

and FINRA are in the process of replacing single-stock circuit breakers with short-term price limits.

In fact, on June 1, 2012, the SEC in Release 34-67091 approved a "limit up-limit down" mechanism that prevents trades in individual stocks from occurring more than a certain amount away from the average price of the security over the preceding 5-minute period. The allowable band is 5% for large-cap stocks and active ETFs, and 10% for most other stocks. If 15 seconds elapse and trading does not resume within the price band, there is a 5-minute trading halt. The new price limits are scheduled to go into effect on February 4, 2013 for a one-year pilot period.

Overall, these price limits seem to be an appropriate speed bump. They are designed to kick in when the potential for adverse selection causes a market failure, so they are economically justifiable. When trading was more centralized, exchanges often exercised the power to initiate a trading halt for exactly these reasons. But in a fragmented market structure, coordination becomes an issue, and a regulatory initiative is an appropriate route to this economically desirable outcome.

8. Recent market glitches associated with HFT

8.A. Knight Capital

Knight Capital Group is one of the largest market-makers in U.S. equities, and it is best known for its arrangements with many brokerage firms to execute orders from retail investors. On August 1, 2012, this high-frequency trading firm introduced a new trading algorithm that was apparently rushed into service without sufficient testing. This rogue algorithm accumulated large positions in 148 NYSE-listed stocks over about 45 minutes at the start of the trading day

before Knight pulled the plug on its trading. Ultimately, including the cost of liquidating its position, Knight incurred losses of $440 million from its trading that morning.

While Knight traded unusually large quantities of the affected stocks and moved share prices significantly, the price moves were typically not fast enough or sharp enough to trip single stock circuit breakers. In addition, single stock circuit breakers do not kick in until 9:45am, 15 minutes after the open. As a result, trading was halted in only five of the stocks involved in the Knight glitch. In the wake of the May 6, 2010 Flash Crash, trade cancellation policies had also been changed, setting a high bar for cancelling trades, specifically so that firms would bear the costs if their orders caused wide temporary price swings. This had the intended effect, as trades were ultimately cancelled in only six smaller stocks that had particularly wide price swings, and Knight was on the hook for nearly all of its unintended trades.

Since this event, the SEC has promised action, but it is not clear that much new regulation is necessary in response to this event. Knight bore the direct cost of its actions, and going forward, this event will remind every market participant of the importance of testing and monitoring its computerized trading. In fact, rogue algorithms should be much less likely now, and firms are much more likely to pull the plug quickly if they observe any unusual trading by their algorithms. However, such errors have the potential to impose externalities on others. For example, this level of trading losses might have bankrupted a smaller firm, imposing losses on clearing firms or other counterparties. Larger trading losses could have systemic effects. Thus, it may be appropriate for exchanges or regulators to mandate a so-called "kill switch" that would quickly shut off a market participant's access and limit any potential spillover effects.

8.B. The Facebook IPO

Earlier this year, Nasdaq had serious computer problems on the first day of trading for Facebook shares. These computer problems appear to be the result of computer software that was not able to handle the pace of order submissions and cancellations by both humans and computer algorithms.

After its initial public offering (IPO) was priced on Thursday, May 17, 2012, Facebook was scheduled to open for trading around 11:00am on the following day. NASDAQ's "IPO Cross" opening process is designed to collect buy and sell orders up to the time of the cross, match as many buyers and sellers as possible at a market-clearing opening price, and allow continuous trading of the stock thereafter. According to Nasdaq, the IPO Cross took a few milliseconds to run and then, as part of its double-checking process, the software looked at the order book again to make sure no new orders had arrived and no existing orders had been cancelled during those few milliseconds. If the order book had changed during the calculation period, the IPO Cross restarted the calculation process. Because of the size and interest level in the Facebook IPO and the presence of algorithms continuing to submit and cancel orders in Facebook shares, the IPO Cross essentially became stuck in an infinite loop.

At 11:30am, Nasdaq opened Facebook trading using a secondary matching engine based on orders present at 11:11am. However, opening cross execution reports were not disseminated until about 2pm that day. Around then, market participants also learned that IPO Cross orders entered or cancelled between 11:11am and 11:30am had not been incorporated into the IPO Cross. While Nasdaq's trading in Facebook was normal for the rest of the afternoon, it is clear that a short-lived software problem caused tens of millions of dollars of losses to investors and their broker-dealers, and it will probably take some time to sort out all of the claims for compensation.

The basic lesson of the Facebook IPO is the same as the lesson of the Knight glitch: critical software needs to be tested thoroughly. Critical software that runs infrequently needs even more testing. As in the Knight case, the incentives are already in place. This event will almost surely cost Nasdaq tens of millions of dollars, and as a result every exchange will now make sure that IPO trading technology (and technological changes more generally) have been thoroughly vetted. Additional regulatory action is probably unnecessary.

9. Other regulatory issues associated with HFT

Regulators in the US and abroad are considering a number of initiatives that can be directly traced to concerns about HFT. However, many of the issues associated with HFT are the same issues that arise in more manual markets. For example, there is concern about the effects of a two-tiered market, where HFT currently has a speed advantage over a second tier of market participants. In a floor-based market, such as the New York Stock Exchange prior to 2005, there were also concerns about a two-tiered market. Within that market structure, the advantages went to those with physical access to the floor, including specialists, floor brokers, and floor-based proprietary traders. To limit those structural advantages, regulators and exchanges instituted rules governing the behavior of floor-based market participants. For example, the NYSE specialist was always last in line at a given price. As another example, most markets imposed severe limits on ability to do proprietary trading while handling customer order flow. Furthermore, there were occasional enforcement cases to rein in abusive behavior. About 10 years ago, several NYSE specialists faced criminal charges due to trading behavior that appeared to disadvantage customers.

Many of those abuses in the floor-based era were due to a lack of perfect competition. Specialists wielded some monopoly power, as did the NYSE. In the current automated market

environment, regulators are largely relying on competition to minimize any abuses that might arise. This can be seen prominently in Regulation NMS, which requires exchanges to provide non-discriminatory access to all comers. Exchanges can charge HFT for co-location services, for example, but they must charge the same amount to any entity receiving the same service.

If there is some sort of market failure, however, then robust competition may not always be the solution, and regulation may be in order. Now under consideration in Washington are a number of potential regulatory initiatives that are directly or indirectly related to HFT. Some of the more important regulatory initiatives are discussed below. In thinking through each one, it is important to confirm that there is indeed a market failure that market participants cannot correct on their own, assess the importance of the market failure, and gauge the likely costs and benefits of the proposed regulatory approach.

9.A. Consolidated order-level audit trails

Robust enforcement of securities laws and exchange rules is important to ensure investor confidence in equity markets. Audit trails have been an important source of data for market surveillance by internal and external regulators, and in many ways, there is nothing new about surveillance of HFT. HFT introduces two new wrinkles into surveillance: their trades may take place on many different venues, and oversight may require a detailed examination of order-level audit trails.

With increased fragmentation in equity trading across trading venues, many trading strategies make use of multiple venues. Thus, regulators may need to obtain order-level and trade-level data from multiple venues and then integrate those data sets together to form a complete picture of the relevant trading activity. At present, all trades and exchange quotes are reported to the consolidated tape, and this is often sufficient for investigators. However, a

high-frequency trading strategy could involve abusive order submission and cancellation behavior, and it could be important to observe contemporaneous order-level data from multiple venues. Right now, integrating order-level data is difficult, because exchanges do not share common data formats for order-level data. Thus, requiring common reporting standards and formats for order-level data would seem to be fairly uncontroversial.

Regulators have suggested that a near real-time consolidated order-level audit trail would be valuable, but exchange operators and other participants have stated that the costs of such a system would be prohibitively high. From an economic point of view, this boils down to a standard cost-benefit analysis. Unfortunately, there does not appear to be much public data available to make these cost-benefit calculations, so it is difficult for an outsider to make an informed assessment.

9.B. Capacity issues and excessive order fees

Capacity has long been an issue for trading venues. For example, the NYSE and AMEX closed on Wednesdays during the second half of 1968 because of a paperwork backlog. Dealing with large amounts of exchange information is also an old problem. For example, before 1900, the Wall Street Journal published the size and price of every NYSE transaction. As trading volume increased, the Journal stopped reporting every trade, instead printing only the daily open, high, low, and last sale price for each stock.

HFT has sharply increased capacity requirements for trading venues. For example, the flash crash of May 6, 2010 revealed that the NYSE did not have sufficient capacity to handle the volume of order submissions and cancellations that afternoon. Other market participants have complained about the sheer volume of data associated with U.S. equity trading. For example, as of the first quarter of 2012, there are an average of about 640 million quotes and 28

million trades per day in the U.S. stock markets appearing on the consolidated tape.[2] This only includes the best bid and offer at each exchange; much more order-level data must be processed in order to maintain an up-to-date view of the limit order book. For example, the Nasdaq equity market processes about 1 billion order messages per day as of June 2010, or approximately 13 gigabytes of order-level data, and Nasdaq is only one of several exchanges with similar levels of activity.

From the point of view of economic theory, the current market structure undoubtedly imposes negative externalities on some market participants, including regulators who must collect and sift through terabytes of data as part of their surveillance function. There are a number of potential ways to ameliorate this problem. For example, many market participants do not need access to every order submission, cancellation, and execution. Exchanges are providing more tools for these participants, such as periodic order book snapshots and data feed subsets.

Some trading venues are also responding to bandwidth and data quantity issues by imposing order cancellation or excess message fees. For example, in 2010, Nasdaq boosted rebates for firms with an orders-to-executions ratio of less than 10. Nasdaq now charges a fee for excessive limit order submissions that are more than 0.20% outside of the national best bid and offer (NBBO). Participants are allowed somewhere between 33 and 100 such orders for every executed trade without charge. Additional submitted orders that are more than 0.20% outside the NBBO are subject to a charge of $0.005 to $0.03 per order.[3]

European equity markets have similar fees in place. For example, NYSE Euronext imposes a surcharge of EUR 0.10 on each order above the 100:1 order-to-trade ratio. Nasdaq

[2] U.S. Consolidated Tape Data, available at http://www.utpplan.com
[3] For more details, see http://www.nasdaqtrader.com/Trader.aspx?id=PriceListTrading2.

OMX Stockholm has an order-to-trade limit of 250. Above that ratio, a surcharge of SEK 0.09 (approximately $0.012) is applied per order. For three months during the summer of 2012, Direct Edge also experimented with a similar policy, reducing liquidity provider rebates by $0.0001 per share for accounts exceeding the 100:1 message-to-trade threshold.[4]

Trading venues must make costly investments in technology infrastructure in order to handle HFT and AT, and these order fees represent a sensible pricing mechanism for recovering some of the costs from those who impose them. In addition, trading venues are under some pressure from their users to limit the amount of order flow information that must be processed. Thus, it seems that trading venues are in the best position to make judgments about these tradeoffs and set their fees accordingly.

At the moment, there are no publicly available studies of the effects of these fees. There do not seem to be any obvious effects on liquidity, but the excessive message charges have been minimal so far and should be fairly easy for most HFT to avoid. However, it is important to note that these fees are designed to be borne by liquidity providers, and if the fees have any bite at all, they should result in reduced liquidity provision. Nasdaq's fees seem particularly well-designed to minimize these adverse effects, because there are no excessive message fees for orders that are submitted within 0.20% of the NBBO (a band that works out to a full 10 cents on a stock with a share price of $50). Still, even the Nasdaq fees could thin out limit order books, as additional depth away from the inside quote may not execute often enough to make that liquidity provision economic. Trading venues and researchers should carefully study the effects of these fee initiatives on liquidity provision and market quality before imposing steeper or broader cancellation fees.

[4] See Direct Edge Trading Notice #12-18 and Trading Notice #12-33.

9.C. Minimum order exposure times

The SEC's 2010 concept release on equity market structure mentions the possibility of a required minimum time-in-force for orders. With a minimum time-in-force, orders could not be canceled within, say, 50 milliseconds of submission. Since it takes approximately 20 milliseconds for signals to travel from electronic trading venues in New Jersey to the west coast of the continental U.S., a minimum time-in-force of about 50 milliseconds could be justified on equal access grounds. More often, a minimum time-in-force is suggested as a way to throw sand in the gears by those who are generally opposed to automated, high-speed markets.

A minimum time-in-force would limit so-called "flickering quotes," and this would provide more certainty to liquidity demanders about the available terms of trade. However, the minimum time-in-force appears to be a particularly blunt, poorly considered tool. A minimum time-in-force would force large changes in equity markets and could severely discourage liquidity provision. Liquidity providers would be granting a trading option to liquidity demanders, and this option would have to be priced into liquidity provision, widening bid-ask spreads by at least the value of the option. Minimum order exposure times would also have an asymmetric impact, affecting liquidity providers but having no effect on liquidity demanders. Since many of the observed benefits from HFT are due to more efficient provision of liquidity, this would seem likely to reverse these market quality gains.

Of course, exchange customers themselves might value longer-lived liquidity. If so, a trading venue could itself put incentives in place to encourage longer-lived limit orders. This seems to be more than a theoretical possibility, as Nasdaq OMX is currently experimenting with minimum order exposure times. On its PHLX exchange (formerly the Philadelphia Stock Exchange), Nasdaq provides a "Minimum Life" order type that cannot be canceled for at least

100 milliseconds. These orders receive a larger rebate if they go onto the book and are later executed. If in fact this order type is adopted by some order submitters, it would be useful to study the effects of this initiative.

Another possible way to limit a potential latency arms race is to conduct frequent batch auctions, perhaps one every 100 milliseconds. All orders that arrive prior to the auction would receive equal treatment in the auction, so there would be almost no incentive to be first with an order. Taiwanese equity markets follow this basic model, and crossing networks have some similar properties. Periodic crossing networks, including ITG's POSIT, generally match buyers and sellers at certain points during the trading day, and trade takes place at the midpoint of the NBBO at the time of the cross. Crossing networks attract a modest amount of trading volume, but they seem to be losing ground to other kinds of dark pools, and if they do not operate continuously, they tend to be relatively infrequent, crossing shares only a few times per day. A U.S. trading venue offering frequent batch auctions would be worthy of study, but at the moment, it does not appear that such a venue is likely to appear.

9.D. Securities transaction taxes

Some policymakers who are skeptical of the value of HFT have proposed a transaction tax on financial instruments as a way of limiting HFT and other "excessive" trading while raising revenue for the government. There have been a number of proposals, differing in some of the details, but an illustrative recent example is the "Let Wall Street Pay for the Restoration of Main Street Act of 2009," a proposal by Sen. Harkin and Rep. DeFazio that would subject most stock market transactions in the United States to a tax equal to 0.25% of the transaction value. Trades in other financial assets would also be subject to tax at varying rates.

A 0.25% transaction tax sounds modest, but it would sharply increase investors' trading costs. If a retail investor wants to trade 100 shares of a stock with a $50 share price, that investor typically pays a commission of less than $10, and pays $1 to $2 in the form of a bid-ask spread. The current SEC Section 31 transaction fee of 0.00224% would add $0.11 to the investor's tab. A 0.25% transaction tax would add an additional $12.50 to that transaction, more than doubling its cost.

For a typical mutual fund, the bid-ask spread and price impact might impose costs of around 5 cents for each $50 share. It might pay an additional penny per share in commissions. A 0.25% transaction tax would mean an extra 12.5 cents per share, more than tripling its trading costs. In addition, transaction taxes will affect the returns of mutual fund investors, even if the investors do not buy or sell their mutual fund shares. Even index funds sometimes have to trade, and in that case index fund investors would bear the tax burden. In fact, the Investment Company Institute (ICI) estimates that a transaction tax would increase the annual expenses incurred by index fund investors by one-third.

A transaction tax would also cause stock prices to fall, because the tax would be assessed on the same share of stock every time it changes hands. The present value of the repeated tax can be quite substantial, and it is capitalized into prices. To quantify the effect, assume for simplicity that all investors hold a mutual fund that trades its portfolio once per year and expects a 6% annual return. In a world with a permanent 0.25% transaction tax, the investors' total return net of the tax would be only 5.75%. Investors would lose about 4.17% (= 0.25% / 6.00%) of their total return each year to the tax. Stock prices would have to fall in order to maintain the same after-tax returns. In this example, investors would drive down stock prices by the same 4.17% (more than 500 points on the Dow Jones Industrial Average at current levels) in order to restore the 6% net total return they require.

Securities transaction taxes would also have a direct negative effect on stock market liquidity. In today's equity markets, bid-ask spreads are one cent for almost every large-cap stock. If market-makers are assessed the transaction tax, they would need to widen their bid-ask spreads to recoup the tax. Continuing the earlier example, this would be 12.5 cents per share under current proposals for a stock with a share price of $50. As a result, bid and ask prices would move further apart. Investors would pay more to buy and sell equities, and it might take more time to find a willing party to take the other side of the trade. Markets would become less liquid, and this illiquidity might worsen the downward pressure on share prices discussed earlier.

Transaction taxes may also worsen stock market volatility. Transaction tax proponents often suggest that current levels of trading activity must be creating excess volatility in stock prices, and this volatility can be reduced by throwing sand in the trading gears via a transaction tax. This is an appealing argument—less trading means less volatility—but it does not match the data. In fact, Jones and Seguin (1997) and others find that transaction taxes increase volatility. The intuition is that if trading is expensive, stock prices must depart further from fundamental value before any market participant has an economic interest in bringing them back into line. Thus, transaction taxes would worsen both volatility and the efficiency of prices.

Given these problems, it is not surprising that other countries have introduced transaction taxes and then repealed them a short time later, once all the effects became apparent. Sweden provides an instructive data point, as analyzed in Umlauf (1993). In 1984, Sweden introduced a 1% transaction tax, and the tax rate doubled two years later. Almost immediately, about 60% of trading volume in the most active Swedish stocks moved from Stockholm to London. The Swedish stock market fell by 5.3% when the tax was initially

announced, increasing the cost of equity capital for Swedish firms. In fact, owing to the stock price declines and the reduced level of trading, Swedish capital gains tax receipts actually fell by more than the amount of transaction tax collected. Thus, in addition to its other negative effects, the transaction tax even failed to raise revenue. Sweden eliminated its transaction tax in 1991, but the country never regained its previous share of trading in Swedish stocks. Much of the trading and the related jobs simply stayed in London.

Overall, the academic evidence indicates that increased transaction taxes will increase volatility, reduce price efficiency, and worsen liquidity. Securities transaction taxes would also increase trading costs and cause trading to move offshore. Most importantly, securities transaction taxes would lower stock prices and increase the cost of equity capital for firms, reducing corporate investment and damaging GDP.

10. Conclusions

Based on the vast majority of the empirical work to date, HFT and automated, competing trading venues have substantially improved market liquidity and reduced trading costs for all investors. Share prices are almost surely higher as a result of this reduction in trading costs, benefiting long-term investors. Higher share prices also have favorable implications for firms' cost of equity capital. With a lower cost of capital, firms are likely to invest more, with commensurate increases in GDP and other measures of economic activity.

In specific terms, HFT has sharply increased competition in market-making, and bid-ask spreads are much narrower as a result. Stock prices are more efficient as a result of HFT activity. Overall, there is no evidence of any adverse effect due to HFT in the average results.

Perhaps the only concern supported by the data is that HFT may not help to stabilize prices during unusually volatile periods. The flash crash of May 6, 2010, when the S&P 500 fell

by almost 10% in the space of less than 15 minutes, was not caused by HFT but by a mutual fund's submission of a rapid sequence of large sell orders during a volatile trading session. Initially, HFT and other intermediaries helped to stabilize prices by buying, but HFT were soon overwhelmed and rapidly liquidated their positions by selling stocks and S&P futures contracts, thereby exacerbating the decline. However, there are many historical cases where intermediaries step aside at times of extreme volatility, so this appears to be a fairly generic feature of equity markets rather than a specific problem with HFT.

A number of regulatory issues are being considered that could affect HFT. Some new policies, such as short-term price limits and circuit breakers, seem well-crafted to address specific problems that arose during the flash crash. Other minor regulatory tweaks such as kill switches may be in order, but those formulating policy should be especially careful not to reverse the liquidity improvements that we have experienced in the U.S. over the past few decades.

References

Baron, Matthew, Jonathan Brogaard, and Andrei Kirilenko (2012), "The trading profits of high-frequency traders," November 2012 working paper, University of Washington.

Benos, Evangelos and Satchit Sagade (2012), "High-frequency trading behaviour and its impact on market quality: evidence from the UK equity market," Bank of England Working Paper No. 469, December 2012.

Boehmer, Ekkehart, Kingsley Fong, and Julie Wu (2012), "International evidence on algorithmic trading," working paper, EDHEC.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang (2012), "Shackling short sellers: the 2008 shorting ban," working paper, EDHEC.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, "Equilibrium high-frequency trading," October 14, 2011 working paper, available at http://ssrn.com/abstract=2024360, retrieved May 31, 2012.

Brogaard, Jonathan (2011a), "The activity of high frequency traders," December 2011 working paper, available at http://ssrn.com/abstract=1938769, retrieved May 29, 2012.

Brogaard, Jonathan (2011b), "High frequency trading and market quality," December 2011 working paper, available at http://ssrn.com/abstract=1970072, retrieved May 29, 2012.

Brogaard, Jonathan (2012), "High frequency trading and volatility," January 2, 2012 working paper, available at http://ssrn.com/abstract=1641387, retrieved May 29, 2012.

Domowitz, Ian (2010), "Take heed the lessons from the 1962 flash crash", June 21, 2010, available at http://www.advancedtrading.com/exchanges/225700888, retrieved May 29, 2012.

Easley, David, Terrence Hendershott, and Tarun Ramadorai (2009), "Levelling the trading field," November 19, 2009 working paper, UC Berkeley.

Easley, David, Marcos M. Lopez de Prado, and Maureen O'Hara (2011), "The microstructure of the 'flash crash': flow toxicity, liquidity crashes and the probability of informed trading," Journal of Portfolio Management 37(2):118-128.

Easley, David, Marcos M. Lopez de Prado, and Maureen O'Hara (2012), "Flow toxicity and liquidity in a high frequency world," Review of Financial Studies 25(5):1457-1493.

Egginton, Jared F., Bonnie F. Van Ness, and Robert A. Van Ness, "Quote stuffing," March 15, 2012 working paper, available at http://ssrn.com/abstract=1958281, retrieved May 30, 2012.

Foucault, Thierry, Johan Hombert, and Ioanid Rosu (2012), "News trading and speed," working paper, HEC.

Gai, Jiading, Chen Yao, and Mao Ye (2012), "The externalities of high-frequency trading," November 16, 2012 working paper, University of Illinois.

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

Hasbrouck, Joel and Gideon Saar (2012), "Low latency trading," working paper, New York University.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld (2011), "Does algorithmic trading improve liquidity?" Journal of Finance 66(1):1-33.

Hendershott, Terrence and Ryan Riordan (2011), "High frequency trading and price discovery," working paper, UC Berkeley.

Hendershott, Terrence and Ryan Riordan (2012), "Algorithmic trading and the market for liquidity," forthcoming, Journal of Financial and Quantitative Analysis.

Jovanovic, Boyan and Albert J. Menkveld (2011), "Middlemen in limit-order markets," October 24, 2011 working paper, available at http://ssrn.com/abstract=1624329, retrieved May 28, 2012.

Jones, Charles M. and Paul J. Seguin (1997), "Transaction costs and price volatility: evidence from commission deregulation," American Economic Review 87(4):728-737.

Kirilenko, Andrei A., Kyle, Albert S., Samadi, Mehrdad and Tuzun, Tugkan (2011), "The flash crash: the impact of high frequency trading on an electronic market," May 26, 2011 working paper, available at http://ssrn.com/abstract=1686004, retrieved May 29, 2012.

Madhavan, Ananth (2012), "Exchange-traded funds, market structure and the flash crash," January 13, 2012 working paper, available at http://ssrn.com/abstract=1932925, retrieved May 29, 2012.

Martinez, Victor Hugo and Ioanid Rosu (2011), "High frequency traders, news and volatility," December 29, 2011 working paper, available at http://ssrn.com/abstract=1859265, retrieved May 31, 2012.

McInish, Thomas H., James Upson, and Robert A. Wood (2012), "The flash crash: trading aggressiveness, liquidity supply, and the impact of intermarket sweep orders," May 23, 2012 working paper, available at http://ssrn.com/abstract=1629402, retrieved May 29, 2012.

Menkveld, Albert J. (2012), "High frequency trading and the new-market makers," February 6, 2012 working paper, available at http://ssrn.com/abstract=1722924, retrieved May 28, 2012.

Moallemi, Ciamac C. and Mehmet Saglam (2011), "The cost of latency", May 27, 2011 working paper, available at http://ssrn.com/abstract=1571935, retrieved May 30, 2012.

Pagnotta, Emiliano and Thomas Philippon (2012), "Competing on speed," April 27, 2012 working paper, available at http://ssrn.com/abstract=1967156, retrieved May 31, 2012.

Riordan, Ryan and Andreas Storkenmaier (2012), "Latency, liquidity and price discovery," forthcoming, Journal of Financial Markets.

Umlauf, Steven R. (1993), "Transaction taxes and the behavior of the Swedish stock market," Journal of Financial Economics 33(2):227-240.

United States Commodities and Futures Trading Commission and Securities and Exchange Commission (2010), "Findings regarding the market events of May 6, 2010," Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues, September 30, 2010.