The economics and regulation of secondary trading markets

Ryan J. Davies *
Erik R. Sirri **

Prepared for the Columbia University New Special Study of the Securities Markets

Abstract

We examine current areas of concern in the regulation of secondary trading markets, including questions concerning market fragmentation, protected orders, minimum tick size, maker-taker pricing models, transparency and dark liquidity, algorithmic and high-frequency trading, the duties and obligations of broker-dealers who handle customer orders, system robustness, flash crashes and episodic liquidity, exchange traded funds, distributed ledger technology, the ownership and usage of market data, and the governance of market data plans. We also discuss potential ways to reduce transaction costs and improve transparency in fixed income markets. To the extent possible, we summarize key findings from the existing academic literature, describe directions for future research, and offer suggestions for future rulemaking.

Keywords: Regulation NMS, Regulation SCI, trade-at rule, dark pools, tick size pilot, high frequency trading, best execution obligations, order type complexity, payment for order flow, maker-taker pricing models, flash crashes, ETFs, designated market makers, quote stuffing, blockchain, market data ownership, fixed income secondary markets

* Babson College, Wellesley, MA 02457, (781) 239-5345, rdavies@babson.edu.
** Babson College, Wellesley, MA 02457, (781) 239-5799, sirri@babson.edu.

We thank Charles Jones, Carole Comerton-Forde, Steven Gordon, and participants at the New Special Study of the Securities Markets Initiating Conference for helpful comments and suggestions.
I Introduction

Capital markets are an essential feature of any well-functioning modern economy. These markets link entities with surplus capital to companies and corporations that can put capital to productive use. Markets establish prices that provide important signals for the efficient allocation of capital among corporations and their projects. In the United States, equity markets form an important segment of domestic capital markets. Stock market capitalization of publicly traded companies in the United States is approaching $30 trillion, about one-third greater than the size of domestic bank assets, with average daily trading volume approaching $300 billion. In the United Kingdom, the stock market capitalization is historically only about one-quarter the size of bank assets, with the ratio being roughly the same for Germany and France.

This paper addresses questions of regulation and design of domestic secondary trading markets. The operation of these markets is subject to comprehensive regulation, primarily by the Securities and Exchange Commission (SEC). Financial Industry Regulatory Authority (FINRA) is the self-regulatory organization (SRO) that oversees broker-dealers, over-the-counter markets, and certain stock exchanges. Many other regulatory organizations also have important roles in trading markets. There are over 40 SROs associated with exchanges and clearing agencies, including specialized SROs such as the Municipal Securities Rulemaking Board (MSRB), which oversees municipal markets. Other regulators that have roles in secondary trading markets include state securities regulators as well as certain federal regulatory bodies such the Federal Reserve Board.

Actual secondary trading markets bear little resemblance to the idealized markets of neoclassical economics. Because trading markets lack complete information, manifest substantial frictions, and exhibit various externalities, it is generally accepted that regulation is
required if they are to effectively perform their basic informational and allocational role. The core framework for the regulation of secondary market trading is embedded in Securities and Exchange Act of 1934.\(^1\) Crucial amendments to this basic body of law occurred with the passage of the Securities Act Amendments of 1975, which inserted Section 11A into the 1934 Act.\(^2\)

Without providing any specific roadmap or requirements, Congress charged the SEC in Section 11A with creating a national market system. The basic goals for this system were: economically efficient execution of transactions; fair competition among broker-dealers, among exchanges, and between exchanges and other markets; ready availability of quotation and transaction information to broker-dealers and investors; the ability of broker-dealers to execute orders in the best market; and an opportunity, consistent with the other goals, for investors to execute orders without the participation of a dealer.

An economically significant change to the core mission of the SEC occurred in 1996 with the signing of the National Securities Markets Improvement Act (NSMIA).\(^3\) Prior to this, the SEC had most often expressed its mission as primarily one of investor protection, with a secondary goal of promoting fair and orderly markets. With the adoption of NSMIA, Congress expressly charge the SEC, when engaged in rulemaking, to also consider “[…] in addition to the protection of investors, whether the action will promote efficiency, competition, and capital formation.” The change is important because of the inclusion of capital formation in the SEC’s mission. Up to this point, the SEC resembled a policeman, ever vigilant for malfeasance, but playing little attention to either the direct cost or the opportunity cost associated with deterring that malfeasance. By mandating the consideration of capital formation and efficiency, the SEC

---

\(^1\) Pub. L. 73–291, 48 Stat. 881
\(^2\) Pub. L. 94–29 (1975)
had to balance its traditional concerns of investor protection with the full economic cost of implementing those policies.

The core principles listed above, as well as the body of rulemaking promulgated over the last 80 years, allows us to extract a series of economic precepts that arise when the SEC, FINRA, and other securities regulators adopt rules governing the participants and institutions of our exchange and over-the-counter markets. While by no means an exhaustive, the discussion below serves to provide an economic structure to understand the types of problems faced by securities regulators in matters related to secondary market trading.

(1) Mandate the production and dissemination of accurate information. Secondary markets manifest significant asymmetric information issues. The value of a firm’s securities depends on its future economic prospects, which change constantly as the firm engages in new ventures and realizes the outcomes of previous investments. Investors, who run the gamut from individuals to sophisticated and large financial institutions, have differential access to such information as well as a differential ability to process it. While little can be done about the difference in processing capability, through innovations like EDGAR the SEC seeks to create a level playing field for core information about public firms and the securities they issue. Even with these and other efforts, asymmetric information remains an important aspect of trading in secondary markets. Not all information that is asymmetric is fundamental, in the sense that it affects firm's cash flows directly. In the case of trading as we shall see below, some of the most important information concerns orders and trades that either have been, or are about to be, public disseminated through facilities of the national market system. Preferential access to that information, even by fractions of a second, can convey a material advantage to traders. Through its regulatory
powers, the SEC seeks, to the extent possible and consistent with its rules, to remove informational disadvantages. Last-sale reporting and the public dissemination of real-time quotes are archetypal examples of this effort.

(2) Alleviate agency problems. Financial markets, and especially trading markets, are fraught with issues of delegation. Investors who wish to deploy their savings in equity and bond markets often retain financial advisors to assist them in making their investment choices. Having made those choices, these same investors cannot on their own find an end counterparty with whom to buy or sell securities. They must retain the services of an intermediary who, in turn, must go to a market center to complete the trade. And even once the trade is completed, investors do not retain ownership and control of their own securities positions. Instead they generally entrust them to their broker-dealer who, in turn, maintains their position at a clearing corporation or central depositary. Each of the above examples gives rise to a potential principal-agent problem, where the principal has retained the agent to perform some critical function associated with secondary market trading. Left to their own devices and freed from any regulatory constraints, such agents may follow courses of actions inconsistent with the wishes and desires of their principal. Specific examples might include a broker who fails to buy a security at the cheapest price available in the market, or a financial advisor who recommends not the best security for an investor, but one that provides the advisor with the largest sales commission. Even intermediaries are subject to agency problems, such as when an introducing broker leaves its customers’ funds on deposit with a clearing broker. Clearing brokers who hold securities on behalf of introducing brokers have an incentive to use those securities for their own business purposes, for example through rehypothecation. SEC rules seek to
mitigate these agency problems where possible. When this cannot be done, regulators often fall back on tools such as disclosure and informed consent.

(3) **Constrain market power.** To be sure, securities markets are competitive. There are approximately 4,000 registered broker-dealer firms, over 20 registered exchanges (equities and options), and over 50 alternative trading systems (ATSs) at which an instrument may be either bought or sold. While for the most part the SEC tries to stay away from price regulation, it does try to foster environments in which market forces can work to ensure competitive outcomes. For example, throughout the 1970s and 80s, the New York Stock Exchange (NYSE) dominated trade in listed equity securities. Though there were six or eight regional exchanges around the country that had the ability to trade NYSE-listed equities on their local markets, their individual market shares were tiny relative to that of the NYSE. The SEC consciously adopted regulatory policies that allowed these exchanges to remain in existence. By doing so, the SEC created a contestable market for exchange services, and the regional exchanges were responsible for some important innovations in trading over the ensuing decades. Examples of how the SEC accomplished this include the practice of preferencing, whereby individual orders could be preferentially routed to regional exchanges.\(^4\) Once there, the orders traded in an environment where they generally did not interact with the bids and offers from off-exchange participants. Another example is the framework used to regulate broker-dealers. The SEC does not adopt a prudential attitude toward broker-dealers. It fosters an environment where entry and exit costs are low, and individual customers can get quality service from even the smallest firms. In part, it does this because of a

---

\(^4\) See Peterson and Sirri (2003) for a description of the practice of preferencing.
regulatory regime that is little concerned with the financial solvency of the broker-dealer, but instead about the safety of customer property. This regime ensures that should a broker-dealer fail, customer property is appropriately segregated and all cash and securities can be returned to customers in short order.

(4) Solve coordination problems. Secondary trading requires the coordination of a host of actors to move securities from order initiation, to trade execution, to confirmation, to clearing, and through settlement. When coordination is absent or insufficient, chaos can result. An oft cited example is the paperwork crisis of the mid-1970s when the exchanges had to shut their doors every Wednesday afternoon simply to get caught up on the processing of paper tickets associated with trades executed over the previous week. The SEC accomplished the required coordination through a myriad of mechanisms. For example, in the United States, market centers are able to jointly synchronize their own clocks to a common time. While the precision of the synchronization may face increasing demands over time, one need only look at attempts in Europe with MiFID to see the costs associated with the lack of such a simple coordination tool. Another more recent example is the requirement for private linkages among equity market centers. These private linkages operate at much higher throughput than the channels provided by traditional gateways through which core market data proceed. Though still a work in process, the SEC has improved the speed with which orders and information can move around the various market centers.

---

6 Markets in Financial Instruments Directive.
7 See Regulation NMS Adopting Release, 70 FR No. 124, at 37497.
(5) *Ensure integrity of systems and infrastructure.* The SEC has a developed program in place to ensure the integrity and robust operation of key market intermediaries such as exchanges, data providers, and clearing agencies. Regulation SCI, which was promulgated in December 2014, is designed to reduce the occurrence of systems issues; improve resiliency when system problems do occur; and enhance oversight and enforcement of securities market technology infrastructure.\(^8\) Other examples include the examination and inspection process carried out by the staff of the Office of Compliance Inspections and Examinations at the SEC and by the field office staff of FINRA. Among other goals, these examinations seek to improve compliance with securities laws and monitor risk-taking by registered entities.

The examples above are merely illustrative, and represent an attempt to classify various activities of regulators in a traditional economic framework. It should be clear, however, that the SEC is a consummately legal body. It is an organization primarily staffed with lawyers, who in turn are charged with administering and enforcing our nation's securities laws. In the words of Christopher Cox, a recent chairman of the SEC, “First and foremost, the SEC is a law enforcement agency [...].”\(^9\)

The remainder of the paper is organized as follows. Section II provides a brief description of the key participants in market centers involved in secondary trading markets. Section III forms the core portion of our paper. It discusses the aspects of secondary trading in equity markets that, in our opinion, are most worthy of additional regulatory attention. These

---

\(^8\) See Regulation Systems Compliance and Integrity, 79 FR 72251.

\(^9\) Testimony of Christopher Cox, Chairman of the U.S. Securities and Exchange Commission before the Committee on Banking, Housing, and Urban Affairs, United States Senate, September 23, 2008.
topics include a discussion of market fragmentation, transparency, algorithmic and high-frequency trading, duties and obligations of brokers, system robustness, and secondary market data. The fourth section discusses secondary trading in fixed income markets. The final section briefly concludes.

II Participants and market centers

There are a diverse group of participants in U.S. equity markets. Over the last 50 years, equity ownership has been concentrating in the hands of institutional investors as they intermediate the market for the savings of retail investors. This intermediation comes in the form of retirement plans, such as 401(k) and 403(b) plans, defined benefit plans (though these have been on the decline), and pooled investment vehicles, such as mutual funds and bank-administered collective investment vehicles. Table 1 shows that the share of equity held by households (including non-profit organizations) has fallen from 43.1% in 2001 to 37.3% in 2015.

In terms of broker-dealer intermediaries, FINRA reports that it has 3,816 registered securities firms in February 2017, which is down from 5,005 a decade earlier in 2007. There are likely a host of reasons for this decline, but certainly one of the reasons is the preference of financial advisors to conduct their retail-facing business using the organizational form of a registered investment advisor rather than an introducing broker. There are a number of benefits to this organizational form, including the ability to easily charge account wrap fees and the absence of an oversight by an SRO. Smaller advisors need not even register with the SEC if their assets under management are low enough.

---

10 https://www.finra.org/newsroom/statistics
Equity trading can occur on any of 12 registered public exchanges, over 30 ATs,\textsuperscript{11} or at off-exchange broker-dealers, including internalizing broker-dealers.\textsuperscript{12} Trade conducted on a public exchange is reported to the consolidated tape, and includes an identifier for the market on which the trade occurred. No such identifier is required for trade on ATs or at off-exchange broker-dealers. SEC and FINRA does require the reporting of these ATs and off-exchange trades, however, and both the NYSE and NASDAQ have created Trade Reporting Facilities (TRFs) for this express purpose. Table 2 provides a snapshot of equity trading volumes across exchanges and the two TRFs for a typical week in 2017. The table does not strictly distinguish between dark and lit order flow, as trade conducted on lit exchanges has the potential to make use of dark orders types. Deutsche Bank reports that the share of high frequency trading appears to have recently plateaued, accounting for approximately 40% of market volume in 2014.\textsuperscript{13}

Table 3 provides a weekly snapshot (week of February 20, 2017) of the trading volumes across the ATs reporting to FINRA. The Tab Group reports that in Q2-2016, equity market volume was split between 56.9% on lit venues and 43.1% on dark venues. Of the dark volume, Tabb Group finds that 51.4% was through retail wholesalers and single-dealer platforms, 30.0% occurred on dark ATs, and 18.6% was hidden exchange volume.\textsuperscript{14}

Retail investors typically trade through introducing broker-dealers of integrated firms such as Merrill Lynch or Morgan Stanley. Broker-dealers with retail customer orders to execute often send these orders to specialized broker-dealers, known as wholesale market makers or “internalizers” in return for a compensating payment. Recent data suggests that vast majority of

\textsuperscript{11} See https://www.sec.gov/foia/docs/atslist.htm for a complete current list of ATs.
\textsuperscript{12} Angel, Harris, and Spatt (2015) provides a nice summary of trends in market quality metrics for equity markets.
\textsuperscript{13} High frequency trading: Reaching the limits. Deutsche Bank Research. May 24, 2016.
\textsuperscript{14} Source: Executive Summary, TABB Equity Digest: Q2-2016, Valerie Bogard, December 8, 2016. Available at: https://research.tabbgroup.com/report/v14-071-tabb-equity-digest-q2-2016
marketable retail orders are sent to these wholesalers. As an example, Scottrade, an introducing
broker that makes use of wholesalers to execute its customer flow, sent a substantial amount of
its order flow to wholesalers, as shown in Table 4. Wholesale market makers typically use
algorithms to determine whether to execute an order, in whole or in part, as a principal or
whether to send it to other trading centers, including exchanges and dark pools.

III Equity market secondary trading

III.A Regulation NMS and market interaction rules

When Regulation NMS was adopted in 2005, it was put in place to address a host of
perceived shortcomings in U.S. equity markets. The traditional listed market, dominated by the
NYSE, was looking increasingly dated in the face of advancing technology. The human
specialist was at the center of this floor-based market, while at the same time the NASDAQ was
operating an order-driven electronic market. Trade-throughs, which are the execution of orders
at prices inferior to the current NBBO, were becoming increasingly prevalent. An SEC staff
study found that 2.4% of trades on NASDAQ occurred at prices inferior to the NBBO.\textsuperscript{15} This
tendency was exacerbated by the poor mechanisms in place to link markets around the country,
such as the Intermarket Trading System (ITS) plan and Unlisted Trading Privileges (UTP) plan.

Also, registered exchanges were prohibited from charging access fees, while ECNs could
and did charge fees up to 0.3 cents per share, and in some cases even higher. These access fees
were not reflected in the disseminated public quotes of the ECNs, causing confusing in the
marketplace. ECNs also had the ability to pay liquidity rebates for placing limit orders into their
order books, which challenged the existing frameworks for brokers’ best execution obligations.

\textsuperscript{15} Regulation NMS Adopting Release, 70 FR No. 124, at 37507 & note 74.
ECNs could also quote in sub-pennies, while the SIP data feeds prevented the exchanges from doing so.\(^{16}\) This allowed the ECNs to step ahead of the exchanges by quoting prices that were better by minimis amounts. Finally, because market data was allocated to exchanges based on the number of trades executed, traders engaged in order shredding, cutting single large orders up into large numbers of small orders, in the hopes of driving more market data revenue to their favored exchange.

Regulation NMS attempted to resolve these problems and restore order to the public markets.\(^{17}\) Though the adopting release for the rule is over 500 pages long, the essence of the regulation is contained in four of its new rules: (i) Rule 603, which allocates market data revenues among market centers to encourage and reward the dissemination of useful trading and quotation data; (ii) Rule 610, which allows private linkages among market centers, and limits access fees to a maximum of three mils ($0.003) per share.; (iii) Rule 611, which protects immediately accessible quotes at automated market centers by requiring incoming orders to interact with the top of their order books; and (iv) Rule 612, which prohibits quoting in less than one-penny increments for stocks priced over one dollar per share.

Not surprisingly, the passage of Regulation NMS engendered a substantial amount of institutional change in secondary trading markets. In this section, we explore the consequences and unresolved issues related to three areas aspects of the regulation: (a) market fragmentation; (b) the order protection rule; and (c) the minimum tick size.

---

\(^{16}\) Securities Information Processors (SIPs) are industry utilities charged with collecting and disseminating quote and trade data. They are discussed in more detail later in the paper.

\(^{17}\) Two of the five sitting SEC Commissioners dissented from the adoption of Regulation NMS because of concerns about its effect on competition and innovation. See Dissent of Commissioners Cynthia A. Glassman and Paul S. Atkins to the Adoption of Regulation NMS, https://www.sec.gov/rules/final/34-51808-dissent.pdf
Market Fragmentation

The SEC’s framework for regulation of secondary equity markets has sought to balance competition among market centers and competition among individual orders. Competition among market centers can lead to innovation and long-term improvements in trading conditions, while competition among orders can lead to greater price discovery and liquidity. In many ways, current regulation promotes innovation, while attempting to create competition among orders through mandated exchange linkages. Consistent with this view, O’Hara and Ye (2011) compare the execution quality and efficiency of stocks with more and less fragmented trading, and conclude that more fragmented stocks have lower transactions costs, faster execution speeds, and greater market efficiency. They conclude that their findings are consistent with U.S. markets being a single virtual market with multiple points of entry.

Issues arise to the extent that linkages between trading venues are not robust or timely. For instance, Rule 611 only protects orders that have been visible for at least one second, and automated trading venues are only required to respond to orders within a one second timeframe. But as trading and quote update latency are now measured in milliseconds (or faster), any delay in processing can cause the linkages between trading centers to be economically severed, leading to effective fragmentation and isolation of trading environments.

---

18 For a survey of the academic market structure literature, see: Equity market structure literature review – Part I: Market fragmentation, Division of Trading and Markets, SEC, October 7, 2013.
19 Jiang, McInish and Upson (2011) argue that market fragmentation allows uninformed traders to segment their order flow to off-exchange venues, allowing a higher concentration of informed trading on lit exchanges, thereby improving price discovery.
20 Aitken, et al. (2017) show that the introduction of Chi-X in Australia led to the arrival of fee-sensitive liquidity providers. Aitken, et al. find that quoted and effective spreads fell as Chi-X market share increased.
21 In the context of ETF trading on the Island ECN, Hendershott and Jones (2005) show that less concentration of trading resulted in weaker competition among liquidity providers, reflecting imperfect competition due to lack of complete transparency and integration.
To better understand the effect of delays in linkages, Bartlett and McCrary (2016) examine the latency of the two SIPs in comparison with those of the exchanges’ direct data feed. Their results show that there is a low likelihood that liquidity-taking trades receive inferior pricing when priced at the SIP NBBO rather than at an NBBO that is constructed by the authors from the private low-latency direct feeds of each exchange. They find, on average, liquidity-taking trades are more likely to find benefit than harm when priced at stale prices appearing in the SIP NBBO.

Growth in the number of trading venues has increased competition. Some of the public debate concerning market fragmentation has been shaped by the winners and losers of this competition, particularly advocates of the legacy exchange systems that have lost order flow. As well, broker-dealers and other market participants have need to adapt to the additional complexity in the market. This complexity has introduced new trading costs, particularly for institutional traders, and may have made markets more vulnerable to large market moves and technological shocks. As well, best execution and the reference price are more difficult to establish in a fragmented market.

Market fragmentation may be a greater concern for small capitalization issuers, for which their already low transaction volume is potentially spread across too many venues. Some market participants have argued that Regulation NMS should allow for more heterogeneity in rules across firms. For instance, NASDAQ has recently argued that removing UTP obligations for

---

22 Ding, Hanna and Hendershott (2014) also study price dislocations between the public NBBO based on consolidated data feeds and prices based on proprietary data feeds. Similar to Bartlett III and McCrary (2016), they find that the brevity of dislocations does not pose meaningful costs for infrequent investors, but that the frequency of dislocations can introduce costs for frequent investors.

23 The debate about market fragmentation was particularly fierce in Europe, as MiFID and then MiFID II, aimed to reduce barriers to competition across national borders. Gomber, et al. (2016) and Davies (2008) describe some of the changes that have resulted.
smaller firms would allow liquidity to be concentrated and would reduce volatility. NASDAQ further argues that removing these constraints would create natural opportunities for other market structures (e.g., batch auctions) to develop.\textsuperscript{24}

Finally, it is important to distinguish between visible fragmentation (dispersal of volume among lit trading venues) and dark fragmentation (dispersal of volume between lit and dark trading venues).\textsuperscript{25} The effects of visible fragmentation can be largely resolved using technology and smart order routing systems. Dark fragmentation, however, can have impacts on price discovery and can cause some orders to be inaccessible to all market participants. In section III.B, we address issues related to dark trading. Some market participants believe that the order protection rule has promoted the growth of dark venues by “constraining the nature of competition on lit venues to factors such as speed, fees, and exotic order types, in contrast to factors that are more appealing to investors, such as liquidity and stability.”\textsuperscript{26} We proceed to discuss this rule next.

Order protection rule

Rule 611 of Regulation NMS, the order protection or “trade-through” rule, is designed to promote intermarket price protection by restricting the execution of trades on one venue at prices that are inferior to publicly displayed quotations on another venue. It is fair to say that the rule has been controversial from the onset. Part of the original controversy was due to the fact that

\begin{footnotes}
\item[25] Degryse et al. (2014) and Gresse (2016) examine both types of fragmentation in the context of European markets.
\end{footnotes}
the rule only applied to quotes accessible for automated execution, which led to the effective end of manual floor markets on the NYSE and a corresponding dramatic market share loss.\textsuperscript{27} Other criticisms of the order protection rule include that it inhibits competition on non-price dimensions, it does not respect time priority across trading venues, and it is difficult to enforce with sub-second trading.\textsuperscript{28}

The order protection rule only protects the displayed top-of-book of each protected market center. At the time of implementation, the decision to protect only the top of a market’s book was based on a belief that protecting the entire book would be technologically infeasible. Given the advances in computing technology, as well as the massive amount of message traffic already occurring, it may be technologically possible today to protect the entire displayed limit order book. However, many market participants would argue against such a change, as critics have already argued that the existing rule overly complicates the system of interconnections among trading venues.\textsuperscript{29}

Closely tied to the order protection rule, is the commonly-used Intermarket Sweep Order (ISO), which allows an institutional trader to access the top of the book across all markets.\textsuperscript{30} Latency between trading venues helps explain some of the usage of the ISO. The ISO allows a trader to release the market center from the time-consuming tasks of checking other trading venues for possible trade-throughs.

\textsuperscript{27} Chung and Chuwonganant (2012) provide empirical evidence that market quality, particularly for institutional traders, decreased subsequent to Regulation NMS, in terms of larger trading costs, greater pricing errors, slower order execution speeds, and lower execution probability. They argue that these results support concerns about the impact of the order protection rule on market liquidity, by reducing the role of NYSE specialists and floor brokers as the liquidity providers of last resort.

\textsuperscript{28} See Blume (2007) for a critical view of the order protection rule, and Regulation NMS more generally.

\textsuperscript{29} For instance, see “Institutions at odds with retail over SEC’s Order Protection Rule” by Rick Baert, Pensions & Investments. http://www.pionline.com/article/20170501/PRINT/305019978/institutions-at-odds-with-retail-oversees-order-protection-rule

\textsuperscript{30} See Chakravarty, et al. (2012) for an overview of ISOs.
Because only the top of the book is protected for ISOs, there may be an incentive to post limit orders on less liquid exchanges where they may be more likely to be at the top of that venue’s book.31 In this manner, the order protection rule, combined with the high usage of ISOs, may be helping to support exchanges that otherwise would be commercially non-viable. Some market participants have suggested that exchanges satisfy a minimum volume threshold to qualify for order protection. They argue that low volume exchanges force them to incur additional costs, such as paying for direct market feeds and managing routing logic.32

The order protection rule allows a trading venue to match the best displayed quote at another venue prior to re-routing the order. Some have argued that this ability discourages displayed liquidity since it allows traders on other venues to trade ahead of existing displayed orders. As well, some market participants are concerned about the growth of dark trading venues. In response to these concerns, a trade-at rule has been proposed as a reform for U.S. markets with the goal of encouraging the public display of orders. In general, a trade-at rule would allow market centers to execute an order against a protected quote up to the amount of its displayed size, with some possible exceptions. Weaver (2014) argues that such a trade-at rule would improve the quality of markets, by dramatically reducing the amount of internalization of order flow.

Closely related to a trade-at rule is a minimum price improvement rule. Foley and Putniņš (2016) examine the impact of implementing minimum price improvement rules in Canada in October 2012 and Australia in May 2013. The rules require that dark trades provide price improvement of one full tick (or half a tick if the spread is at one tick). They show that the

31 It is also possible that market participants may be using exchanges with inverted, taker-maker pricing to extract information from these orders.
effect of the rules is different for one-sided dark trading (i.e., midpoint crossing networks) and two-sided dark trading (i.e., markets with fractional price improvement). In both markets, the rule reduced dark trading, as expected. They show that the reduction in two-sided dark trading resulted in higher quoted, effective, and realized spreads, and lower information efficiency.\(^3\)

The decrease in trading on midpoint crossing networks did not impact market quality. These findings reinforce the notion that dark traders are not a homogeneous group and that a one-size-fits-all regulatory approach may not be effective. As such, any restrictions on off-market trading, such as a trade-at rule, might need to be applied differentially across trading venues. Comerton-Forde, Malinova, and Park (2017), discussed in more detail below, show that minimum price improvement rules in Canada did not benefit all market participants.

The trade-at rule is being tested in U.S. markets as part of the tick size pilot discussed in the next section. In the pilot, certain exceptions to the trade-at rule are provided for block size orders and for conditions that mirror those already covered by the order protection rule (e.g., crossed markets and orders marked Trade-at-ISO). This pilot is unlikely to provide a definitive answer on the effectiveness of the trade-at rule since it is constrained to small capitalization stocks and it is being implemented in conjunction with an increase in the tick size.

The trade-at rule being tested in the tick size pilot is based on posted prices. These prices do not include market access fees and trading rebates, and as such, do not reflect the true net cost of the transaction. Two possible modifications to the trade-at rule could address this problem. One possible modification is to apply the trade-at rule only to posted orders that have access fees at or below a certain level, which could be zero. Alternatively, the trade-at rule could be

changed to apply to prices after including fees and rebates, but this change would require a major change in Regulation NMS and the manner in which prices are quoted and transmitted across venues.

**Minimum Tick Size**

Rule 612 of Regulation NMS specifies the minimum pricing increments for NMS stocks. The rule prohibits market participants from accepting, ranking, or displaying orders, quotations, or indications of interest in a pricing increment smaller than a cent for stocks with prices exceeding $1.00. Some market participants have argued that decimalization has contributed to poor liquidity in small capitalization stocks, and thereby has led to a loss of aftermarket support for new issues and a dramatic decline in new IPOs. The 2012 JOBS Act directed the SEC to run an experiment on the impact of increasing the tick size. In response, the SEC developed the tick size pilot program, which is designed to examine the impact of the minimum quoting and trading increment on the liquidity and trading of small capitalization stocks.

The pilot began on October 3, 2016 and will run for two years. It focuses on companies with market capitalizations less than $3 billion. The pilot has a control group (approximately 1400 stocks), and three test groups (approximately 400 stocks each): the first test group will be quoted at $0.05 increments, but allowed to trade at $0.01 increments; the second test group will be quoted and trade at $0.05 increments, with some exceptions; and the third test group will be quoted and trade at $0.05 increments, as well as subject to a trade-at requirement.

---

34 Jones and Lipson (2001), Goldstein and Kavajecz (2000), and Bessembinder (2003) study the impact of decimalization on trading costs.
35 See, for instance, Weild and Kim (2010).
36 Jumpstart Our Business Startups Act.
The Capital Markets Cooperative Research Centre (CMCRC) has developed a dashboard that monitors the impact of the pilot study on a day-to-day basis. Preliminary results show that a wider tick size has led to a more stable quote, with more exchanges quoting at the NBBO for longer periods, thereby making it easier to trade in size. Effective spreads have widened, and market share has shifted from maker-taker venues to inverted (taker-maker) exchanges. The overall welfare effects from changing the minimum tick size have not been fully quantified yet.

The shift to inverted exchanges from widening the minimum tick size appears to be consistent with Yao and Ye (2015). They argue that artificial price constraints caused by the minimum tick size give rise to speed competition, and reduce the ability to compete on price. They argue that the rule has been a factor in driving high-frequency trading and the proliferation of inverted taker-maker markets. Yao and Ye believe that a binding $0.01 tick size led the usage of inverted taker-maker markets; clearly, the $0.05 tick size in the tick size pilot will be even more binding and is likely to increase usage of these venues.

Kwan, Masulis, and McInish (2015) provide further evidence of the effects of the minimum tick size. They find a discontinuity in the market share of dark ECNs around the $1.00 price threshold, suggesting that the minimum pricing increment rule provides a competitive advantage to these dark ECNs. They conjecture that when spreads are constrained on major exchanges, traders use dark ECNs to enable them to jump the queue of existing displayed limit orders, reducing their risk of delayed execution.

Harris (2013) shows that maker-taker and taker-maker pricing models provide a means for exchanges to provide net quotes in sub-penny increments, thereby undermining the

---

38 The dashboard can be found here: https://www.mqdashboard.com/tick_size#search/nasdaq,nyse/1,2,4/2016-06-01/2017-02-27/false/false/none/false
prohibition on sub-penny quotation pricing in Regulation NMS and allowing certain traders to jump ahead of others. Harris argues that these pricing models produce an agency problem between brokers and their clients, since the broker is incentivized to send limit orders to maker-taker exchanges to earn liquidity rebates and avoid access fees. These agency conflicts are discussed further in the discussion on best execution below. Harris also argues that maker-taker pricing models introduce unnecessary complexity and reduce the transparency of bid-ask spreads for retail investors.

Comerton-Forde, Grégoire, and Zhong (2017) use the tick size pilot as an exogenous shock to the share of trading on inverted taker-maker fee markets. Trading on venues with inverted pricing models increased following the introduction of the trade-at rule, particularly for stocks that were tick size constrained. They argue that the finer price grid made possible by venues with inverted pricing models encourages competition between liquidity providers and improves market quality, with more orders executed and less orders canceled.

In summary, the effective tick size can be larger or smaller than the minimum tick size because of the presence of maker-taker pricing models and other forms of access fees and liquidity rebates, and payment for order flow arrangements. These fees and rebates are not included in quotations, and confuse the measurement of best execution. To reduce potential conflicts of interest, broker-dealers should be required to disclose how access fees and liquidity rebates affect order routing practices and transaction costs.

Importantly, Rule 612 allowed sub-penny trading under two exceptions: mid-quote executions and execution at a price determined through a VWAP algorithm. These exceptions appear to be used frequently by high frequency traders through dark pools. Buti, Rindi, Wen, and Werner (2013) find that approximately 10% of share volume executes at sub-penny
increments. They argue that the ability to undercut existing displayed limit / liquidity orders by trivial amounts, can lead to less passive displayed orders, less depth and larger spreads. Bartlett and McCrary (2015) use a market discontinuity to show that increasing the incentive to use the exception to the sub-penny quote rule increases the rate of trading at the midpoint of the NBBO on dark venues, thereby offering liquidity takers price improvement equal to the quoted half-spread. They argue that this evidence goes against the belief that sub-penny trading offers little or no price improvement.

For larger, more active stocks, it is possible that the tick size is no longer relevant in a trading environment in which quotes are only a probabilistic indication of the likely price of a market order. In an environment with flickering and fleeting limit orders, the latency between order submission and order arrival has the effect that posted quotes are only an input into a probabilistic view of the likely execution price.

While the verdict on the SEC Tick Size Pilot is still out, the SEC has announced plans to run an Access Fee Pilot to examine the effect of maker-taker pricing. As pilot studies are more commonly used, we note that the advantages of using a pilot study to obtain information via a controlled experiment must be balanced against the costs such a study imposes on market participants. It is possible that the pilot study period may not be long enough to establish the end state or equilibrium that would arise from a permanent rule change. Market participants, knowing that pilot will end after a certain period of time, may elect to not devote resources towards developing an optimal reaction to the rule change. Rather than develop a process for handling the pilot study rule, these firms may simply elect to avoid certain trades in the pilot study securities. Given the costs of running a pilot study, it is important to determine upfront the goals of the study and set clear guidelines on what it is trying to measure. The pilot study must
be well designed, such that it can deliver results that are informative and which could not be obtained through less costly, alternative means.

III.B Transparency and dark pools

Dark liquidity and dark trading have always existed on U.S. equity markets. In prior years, NYSE floor brokers were a source of dark liquidity, either leaving large customer orders with the specialist (passive participation) or working them over time as a member of the trading crowd (active participation). Other floor traders were only partially aware of the magnitude of these orders, and they were not visible to the public outside of the floor. Similarly, dark trading also occurred in the so-called upstairs market as an intermediary or broker searched across brokerage firms to locate a counterparty for a large block trade before sending it to the downstairs market for execution. While some information leakage occurred as the order was shopped, most market participants were unaware of these block trades and were not given the opportunity to participate in the trades.

Today, dark liquidity can be found in the various types of hidden or non-displayed orders available on almost all exchanges. These orders allow traders to hide all, or a portion of, their orders on the book, typically at the cost of lost priority to displayed orders at a given price. These hidden orders constitute an important source of dark liquidity. Non-displayed order types are the most commonly used order types on exchanges, constituting more than 25% of orders on BATS and more than 30% of orders on Nasdaq OMX by a recent estimate. SEC market data

40 See Keim and Madhavan (1996) for details.
41 In addition to hidden orders, another source of dark liquidity is institutional orders that have been sent to agency brokers, such as ITG Inc., but have not yet fully revealed to the market.
show that these order types may account for as much as 11% to 14% of exchange-based volume. Interestingly, Bloomfield, et al. (2015) use a laboratory experiment to investigate the impact of allowing traders to hide their orders and they find little effect on overall market outcomes.

Equity trading can occur on traditional exchanges (lit markets) and off-exchange venues (dark markets), such as dark pools, crossing networks, and retail internalizers. In most contexts, dark trading has virtually no pre-trade transparency, but has some post-trade transparency (which may be incomplete, possibly not indicating the venue that executed the trade, and may occur with a lag). Garvey, Huang, and Wu (2014) use proprietary data from a direct market access broker to investigate the reasons that orders are sent to dark markets. In their sample, they find that more sophisticated traders tend to participate more in dark markets, particularly when market conditions in lit markets are challenging (wider spreads and higher volatility). They find that more than 80% of dark orders are executed at a price better than the best price available in lit markets at the time of order submission.

Despite having no pre-trade transparency, in some cases the information from dark trading may also be partially reflected in lit market prices. For instance, Nimalendran and Ray (2013) use a proprietary dataset of transactions on a crossing network to examine whether information leaks from trading on a dark venue to the trading on the lit venue. They show that buyer-initiated trades on the crossing network are followed by more net buy signed trades on the lit market, and vice versa. This pattern is suggestive of concurrent informed trading on both dark and lit venues, which means that the potential negative impact of dark pool trading on price discovery might be less than expected.

In the Australian context, Comerton-Forde and Putniņš (2015) find that high levels of dark trading impede price discovery and cause prices to become less informationally efficient.
They find that dark trading increases adverse selection risk, bid-ask spreads, and price impact on the transparent exchange. The impact of dark trading on price discovery appears to be driven by smaller trades occurring in the dark, rather than offsetting block trades. While the levels in the Australian market may not apply universally, their findings suggest that there may be a tipping point in U.S. equity markets, and it is important to understand where the tipping point is.

Theoretical work suggests that determining the tipping point is not straightforward, and may differ across securities and market conditions. Zhu (2014) develops a model with asymmetric information about the asset value that shows dark pools concentrate price-relevant information on the exchange, improving price discovery but reducing liquidity. Ye (2016) provides a model that shows that the impact of dark pools of price discovery depends critically on traders’ information precision. In her model, dark pools have an amplification effect: price discovery is enhanced by dark trading when information precision is high, and impaired when information precision is low.

When discussing dark pools and their effect on market quality, one should be careful to note that dark pool venues are not all the same. Some dark pools (e.g., Liquidnet Negotiated, Liquidnet H20, Barclays DirectEx) resemble the old upstairs market with infrequent, large trades (often at negotiated prices or crossed at the NBBO midpoint), while other dark pools automate the execution of a large number of tiny trades, often used by market makers as retail internalization pools (e.g., Goldman Sachs Sigma-X, KCG Matchit).

It might be tempting to promote dark trading of large blocks, while discouraging venues that execute smaller orders, which may or may not interact directly or indirectly with retail orders. However, the latter venues play an important role in the marketplace as institutions often

---

43 See Mittal (2008) for a categorization of dark pools.
find it difficult to locate block counterparties of similar size. Menkveld, Yueshen, and Zhu (2017) show that the diversity of dark pools provides a valuable range of options for investors facing a tradeoff between price impact and execution uncertainty. They show that as trading needs become more urgent, investors move from low-cost, low immediacy venues to high-cost, high immediacy venues.

Dark trading, in all forms, serves a critical role. Dark pools provide liquidity for orders that would be too large to send to the market fully displayed. Dark pools often advertise that they allow institutions to trade large positions with reduced market impact, lower information leakage, better pricing, and anonymity. On a trade-by-trade basis, these claims may be correct, but it is difficult to assess how dark trading affects overall market quality as there is an endogeneity issue that cannot be easily resolved. To illustrate, consider the impact of a dark pool on bid-ask spreads. On the one hand, wider spreads on lit markets might cause more order flow to be sent to the dark pool. On the other hand, the decision to send order the order to the dark pool could lead to wider spreads on the lit market by increasing adverse selection risk. Controlled experiments, such as the tick size pilot combined with the trade-at rule, might help with the identification aspect of this issue.

Importantly, the overall impact of dark trading on market quality cannot be captured solely by examining the percentage of trading volume on lit and dark venues. The characteristics of the trades on lit and dark venues matter. For instance, it is possible that lit markets are becoming dominated by short-term trading, while dark markets are being used as the venue of

---

44 Cheridito and Sepin (2014) show in simulation results that the presence of a dark pool lowers the implementation cost of acquiring a large position.
45 Buti, et al. (2011) find that increased dark pool activity improves market quality measures such as spread, depths, and short-term volatility, but that the impact on price-efficiency is more complex.
choice for long-term, informed investors. If the majority of informed trading occurs on dark venues, it will likely have a detrimental impact on price discovery.

The opaque nature of some of the automated dark pools is a potential for concern. For these dark pools, is unclear what algorithms are being used to match trades and whether the operators, or a subset of participants, in the pools are able to extract unfair advantages. Recently, Barclays Capital Inc. paid $70 million in fines for, among other things, failing to properly monitor predatory trading on in their LX dark pool, and Credit Suisse paid $84.3 million in fines for, among other things, failing to identify opportunistic traders and executing 117 million illegal sub-penny orders in their Crossfinder dark pool. The SEC may propose a new rule that would require dark pools to publically disclose more information about their procedures and whether some traders receive preferential access to certain functionality. In addition to the new proposed rule, more complete post-trade transparency, in particular more information regarding the counterparties to all of the trades, may help identify similar issues earlier.

Another form of dark trading is retail internalization. Retail internalization refers to the practice by which marketable retail orders in U.S. equity markets are typically routed to wholesale market makers, rather than the exchange. The wholesale market makers pay the retail brokers making these routing decisions a payment for this order flow. The sector is dominated by five wholesale market makers: Citadel Securities (Citadel Execution Services), KCG

---

Americas LLC, G1 Execution Services, UBS Securities, and Two Sigma Securities. At least 10% of consolidated U.S. equity market volumes is in the form of retail internalization.\textsuperscript{48}

Recently, some jurisdictions have attempted to restrict dark trading and retail internalization. Comerton-Forde, Malinova, and Park (2017) examine the impact of a rule change in Canada requiring dark venues to provide a minimum price improvement (similar to the trade-at rule proposed by the tick size pilot), which effectively ended intermediation of retail orders in the dark. After the rule change, retail orders were sent primarily to the lit market with the lowest fees for marketable orders, improving liquidity with larger displayed depth and tighter spreads. While liquidity on the lit markets improved, retail traders received less price improvement and institutions had higher implementation shortfall costs. Retail brokers paid higher exchange fees, while high frequency market makers captured larger exchange rebates.

In Europe, when MiFID II rules come into effect in January 2018, each dark pool will be limited to trade no more than four percent of the overall trading volume in an individual security, and total dark trading will be restricted to eight percent of overall trading volume.\textsuperscript{49} Based on current dark trading volume statistics, these limits will be binding. Regulators in U.S. markets should watch these developments carefully.


III.C Algorithmic and high frequency trading

While there have been many changes to equity markets over the last 30 years, one of the most important is the transition of trading from a manual to an automated process. This transition has affected not only exchanges and market centers, but brokers and investors as well. Far from being a technical change or merely a step along the way of market evolution, automated trading has reshaped our markets and is almost certainly a permanent fixture of the trading scene.

Traditionally, orders were handled manually from their origination at the desk of a retail or institutional investor to their execution on the floor of an exchange. Over time, technology crept into this chain through devices like the fax machine and later through fixtures of exchanges such as ITS and exchange-based automated execution facilities. The advent of algorithmic trading brought automation to brokers and public traders for the generation and submission of orders.

The nomenclature in this area is confused, and there is no specific definition of what constitutes computerized, algorithmic, or high frequency trading. Elements of computerized trading have been with us for many years. For example, program trading, defined as the simultaneous submission of orders in 15 or more equities, has used automation for years. And clearly computers have been used in the investment process for decades. Only more recently has automation come to the generation and management of orders on a continuous basis.

While there are no precise definitions, for our purposes we define algorithmic trading to be the use of computer-based algorithms to generate, submit, and manage child orders derived from a larger parent order. Examples of such algorithms listed by a major broker-dealer on its

---

50 An early example of such an exchange facility was SOES, NASDAQ's small order execution system. ITS was an electronic system that linked exchanges and allowed exchange members to execute orders in any market with the best price, not just the one for which they were a member.
website include algorithms based on the VWAP (volume weighted average price), TWAP (time weighted average price), volume in line, price in line, pairs, implementation shortfall, liquidity-seeking strategies, float, hidden DMA, open and close strategies, and bespoke strategies.\(^\text{51}\) Of course, in many ways algorithmic trading merely automates a strategy that previously existed in manual form, making it more efficient and customizable.

*High frequency trading* (HFT) is more difficult to define. The SEC defined the attributes of HFT to include the use of extraordinarily high speed and sophisticated programs for generating, routing, and executing orders; use of co-location services and individual data feeds offered by exchanges and others to minimize network and other latencies; very short time-frames for establishing and liquidating positions; submission of numerous orders that are cancelled shortly after submission; ending the trading day in as close to a flat position as possible.\(^\text{52}\) As an illustration of the short-term nature of the HFT business, Menkveld (2013) has found that high frequency traders profit only on those positions that are held for less than five seconds.

The speed associated with high frequency trading is truly mind-boggling. Latencies associated with HFT are now well under one millisecond. Data produced by the SEC show that order interaction times can be as low as 50 microseconds. That is, once an order is placed on the books of an exchange, it can either be traded against or canceled, in whole or in part, within 50 millionths of a second. Such rapid speeds are only achieved by using ultrafast hardware throughout the trading process. Even the distance between a broker’s offices and the exchange’s matching engine has become a binding constraint on trading activity. Exchanges now offer colocation services that house traders’ analytic hardware in close proximity to the exchange.


matching engine. The services are available for a fee, but SEC rules require them to be offered on a nondiscriminatory basis. Brogaard, Hagströmer, Nordén, and Riordan (2015) examine a colocation upgrade in Stockholm and find that it improves liquidity for the entire market. The reason cited by the authors is that the market making entities who take advantage of the speed upgrade use the advantage to reduce their losses to adverse selection, allowing them to quote tighter markets for all.

High frequency traders generate a tremendous amount of order traffic to accomplish the execution of a single order. Data show the typical trade-to-order submission ratios are between 2% and 4% on the major exchanges. That is, between 25 and 50 orders are generated for every execution. These submission ratios are even lower for exchange traded products such as ETFs, running well under 1%. The lifetime of these orders can be very short as the governing algorithms implementing their designated strategies by continuously canceling and replacing orders. For example, about 8% of orders are fully canceled in 500 microseconds, and almost half of orders are canceled in less than a second. The high number of orders and their attendant cancellations associated with completing a trade is characteristic of the algorithms used by high frequency traders.

The last ten years has seen a wealth of academic research on HFT. It is beyond the scope of this article to summarize all this research here. Jones (2013), Menkveld (2016), and O’Hara (2015) provide reviews of research touching upon algorithmic and high frequency trading and

---

the issues attendant on the practice. While the articles discussed in these three papers cover a broad range of topics, many of them focus on issues such as information aggregation, pricing efficiency, and the effectiveness of liquidity provision by high frequency traders.

For our purposes, HFT poses some difficulties that must be examined and understood. The bulk of today’s regulatory framework was developed long before the advent of computerized trading, let alone HFT. HFT is not just doing the same thing faster, it fundamentally changes the nature of trading and the way participants in the trading process interact with markets and competitors.

HFT and Liquidity

The issue of perhaps greatest concern to regulators is the provision of liquidity to markets. Traditional liquidity providers were exchange market-makers and specialists. These were organizations that held themselves out as providing a liquid and continuous market in the stocks they covered. They received certain regulator-conferred benefits of time and place in the trading process in exchange for both affirmative and negative obligations in their market-making activities. The most well-known of these liquidity providers was the NYSE’s specialist, who ran a post on the floor of the exchange and acted as an agent for the book of limit orders left with him, receiving a commission for doing so. Liquidity was also provided by the “upstairs” market, a network of block traders and positioners at major brokerage houses who used proprietary capital to acquire and dispose of large blocks of stocks, profiting on the difference between the purchase and sale prices. Finally, there was also an agency search business in which agency brokers would, for a commission, search for counterparties to large block trades.
The advent of Regulation NMS, and especially its framework of allowing private linkages between exchanges, has led to the replacement of traditional market makers by computer-based and high frequency market-making. These opportunistic market makers are not charged with the same obligations of traditional exchange market makers and specialists. For example, they are not required to continuously quote on both sides of the market, nor are they required to fill gaps in limit order books. This led to a justifiable concern that HFT market makers provide liquidity only episodically and when it is in their economic interest to do so. Under certain conditions, this could lead to highly illiquid markets or the short flash crashes that we have observed. That said, little is known about the algorithms used by these new market-making firms. Algorithms are the core intellectual property possessed by HFT market makers. As such, they zealously guard the confidentiality of this software.

The academic literature, however, has provided some important results for questions associated with algorithmic and high frequency trading. For the most part, this literature has generally established that algorithmic and high frequency trading provides benefits to the marketplace. For example, Hendershott, Jones, and Menkveld (2011) examined the introduction of an automated quotation system on the NYSE and its effect on algorithmic trading. This change allows algorithmic traders to effectively make markets via the electronic submission and cancellation of liquidity-providing limit orders. The authors find that algorithmic trading reduced the cost of trading through the narrowing of spreads, which result from a decrease in adverse selection. Algorithmic trading caused price discovery to occur without trading, causing quotes to be more informative and improving pricing efficiency.55

---

55 Notably, these results are found only in large cap and not small cap stocks.
Boehmer, Fong, and Wu (2015) study ten years of algorithmic trading data across 42 international equity markets. They find that algorithmic trading improves informational efficiency and liquidity, but they also find an accompanying increase in volatility. The authors also find that algorithmic trading harms the market quality of the smallest firms, decreasing their liquidity and increasing their volatility. Hendershott and Riordan (2013) use data from the Deutsche Börse that flags orders from algorithmic traders. They find that algorithmic traders primarily submit smaller-sized orders, and provide liquidity through limit orders, thereby narrowing bid-ask spreads when they are wider than average. Algorithmic traders will take liquidity when spreads are sufficiently narrow. Thus, algorithmic traders have the effect of inter-temporally smoothing liquidity.

A number of other papers support the proposition that HFT improves pricing efficiency and decrease adverse selection. Brogaard, Hendershott, and Riordan (2014) find that high frequency traders improve market efficiency through their marketable, as opposed to their nonmarketable orders. These traders have short light information whose use cause trading in the direction of permanent price changes and against short-term pricing errors. The authors find that high frequency traders use nonmarketable orders generally lose to the other informed orders, but these losses are more than offset by spread profits and liquidity rebates. Menkveld (2013) examines trading by a single large high frequency trader across two different European equity markets. He finds that most trades by the trader are placed passively, making money on the spread but losing money on inventory positions held longer than five seconds. Interestingly, the author observes that because central clearing system of the U.S. allows positions acquired on different markets and is to be netted, thereby reducing capital charges, this may explain why U.S. stock markets are more fragmented than those of Europe.
Brogaard, Hendershott, and Riordan (2016) offer a view on how markets may be changing as HFT increases in importance. Using Canadian regulatory data for 15 TSX stocks, the authors show that high frequency traders are responsible for 60% to 80% of price discovery. Most notably, the authors show that high frequency traders react to the trading of others primarily through their use limit orders, not marketable orders. This refutes the notion that high frequency traders behave in a predatory manner by selectively preempting non-HFT trades. However, the authors note that high frequency traders do trade in such a way as to move prices against large non-HFT orders before they can be completed. Of course, this not necessarily indicative of illegal “front-running” behavior, as some have contended, and may result from high frequency traders extracting information from sequential partial executions of large orders.

Shkilko and Sokolov (2016) conduct an interesting analysis that refutes some of the empirical findings above. They examine instances of bad weather disrupting microwave trading networks and reducing the speed advantages of ultralow latency traders. When this occurs, adverse selection falls, trading costs decline, and liquidity improves. Interestingly the authors show that their results depend on tick size. During periods of bad weather, latent liquidity emerges to narrow spreads when tick sizes are not binding. When tick sizes are binding, latent liquidity improves quoted depth. The authors also show that with respect to the use of limit orders by informed traders, when tick sizes bind and queue lengths are long, these traders switch to liquidity-taking marketable orders.

The empirical academic literature is generally consistent with the view that HFT is beneficial to liquidity and pricing efficiency. Though the results are not without exceptions, they are surprisingly consistent across studies and markets. While this is a welcome result from a
regulatory perspective, other issues remain. Below we touch upon some of the additional areas of concern with respect to HFT.\textsuperscript{56}

\textbf{Arms race and fairness considerations}

Market observers have recently become concerned about an arms race developing among cutting-edge high frequency trading operations. These ultralow latency traders seek to get any advantage they can in the contest to be first to the market to execute or cancel an order. Examples include the 825-mile fiber optic cable installed by Spread Networks between New York and Chicago, at a cost of over $300 million. Fiber-optic linkage already existed between these two cities, and the purpose of this cable was to find the shortest possible path, and hence the least time of travel, between the two cities. Several years later, Anova Technologies, in a game of one-upmanship, bettered this technology by installing yet-faster microwave and millimeter-wave lasers to communicate between these two cities, bypassing optical cable entirely. Hibernia Networks, a global telecommunications firm, has created the Hibernia Express, the 4600-km transatlantic fiber optic cable designed to carry trading and other financial data between market centers in the U.S., London, and Europe.

Exchanges themselves do everything they can to maximize their speed of throughput and minimize their own latency. For example, NASDAQ’s matching engine is capable of a sustained 100,000 orders per second at a matching latency of less than 40 microseconds. BATS’ exchange-level latency is less than 200 microseconds, which is the time it takes for BATS to accept, process, and fill a member order, as measured from outside the BATS firewall. In

\textsuperscript{56} Flash crashes, whose causes can also be traced, in part, to high frequency trading, are treated separately in Section III.E.
addition, BATS can process over 400,000 peak messages per second. Contrast that to execution times of 20 years ago were executions were measured in seconds or even minutes.

Today exchanges use high-speed private intermarket linkages, but until as recently as 2007 exchanges communicated via the ITS. The ITS Plan restricted the ability of exchange and NASD (now FINRA) members to trade-through the better-priced quotations of other markets. Sometimes described as “two tin cans and a string,” this was a linkage mechanism that transmitted quote commitments for a trade. Any receiving exchange was allowed at least 30 seconds to respond to a trading commitment, which in turn could not be canceled during those 30 seconds. That meant that anyone using the ITS granted the receiving trade exchange a 30-second option to trade against their order. Put in place to ensure that regional exchanges were integrated with the primary markets on the East Coast, it should not be surprising that such a system was very unpopular with remote traders.

Today the speed of light has become the binding constraint for many aspects of intermarket communication. This is reflected in the rise of co-location services, whereby exchange seeks to locate their members’ servers that house proprietary trading execution systems as close as possible to the matching engine of the exchange. Exchanges charge a tiered fee for doing so, with servers having the closest proximity to the matching engine paying the steepest co-location fees.

In many respects, there is nothing new about this race for greater speed of market access. In the 19th century, a member of the Rothschild family was alleged to have earned a small fortune in London stock market by trading on advanced knowledge of the outcome of the battle of Waterloo. More recently, telegraph, and then telephone, and then handheld electronic

57 See, for example the discussion in Glode, et al. (2011).
communication, has improved access and speed to the floor from outside the exchange. Yet none of these innovations have until now bumped up against the physical limit of communication. Low latency traders desire for the fastest possible access to market centers stems from their desire to be first in the queue for execution. It is doubtful that there are any social benefits the rise from the incremental efficiency of prices over millisecond horizons, and the behavior seen by many as wasteful.

Per-unit profits of high frequency traders are small as their flow-based business is fundamentally different than the traditional equity trading businesses of decades ago. Brogaard, et al. (2014) estimate revenues for HFT at about $.04 per $1,000 traded, and cite to public financial data documenting expenses of about two-thirds of trading revenues. The total profitability of the operation arises from the scale and scalability of large high frequency trading operations. The authors document that high frequency traders only make profits on the larger stocks they trade.

Various suggestions have been made to temper this arms race. Budish, Cramton, and Shim (2015) argue that a solution can be found in periodic batch auctions. The authors suggest that continuous markets be replace be periodic double auctions, occurring as often as every 100 milliseconds. They argue that short-lived arbitrage opportunities arise from market design and limited processing opportunities, leading to a socially wasteful arms race for speed advantages. By batching auctions, speed advantages would be negated as the first-come first-served characteristic of continuous markets would no longer characterize the allocations of batch auctions. While interesting as a theoretical construction, integrating such a platform into the national market system as a registered exchange would require a substantial rewrite of the rules. Among other consideration, Rule 600 of Regulation NMS states that the quotes of a public
exchange are afforded protected status only as long as they are “immediately and automatically” accessible. NMS Rule 611 further requires that trades be executed against the best protected quote, which would likely be problematic for an exchange but pauses to aggregate trades every 100 milliseconds.

A second solution that has been proposed, and in some instances implemented, is the “speed bump,” an intentional delay that slows down access and messaging to the market center. The most well-known instance of this occurs on the new U.S. exchange IEX, which was previously organized as an ATS. IEX creates a 350-microsecond delay by running all external communications through a coil of fiber-optic cable. By creating this delay, IEX argues they have created a market ill-suited to the latency arbitrage strategies of high frequency traders and have effectively created a “safe space” for uninformed non-HFT market participants who wish to avoid being adversely selected against. IEX and its founder, Brad Katsuyama, were the protagonists of Michael Lewis’s popular book Flash Boys when the market center was still an ATS. Since becoming an exchange on June 17, 2016, IEX’s market share has remained low at approximately 2.2%. So while the speed bump maybe a potential solution to the wasteful arms race, it is too soon to draw any definitive conclusions about its commercial acceptance by the marketplace.

Perhaps more concerning than the social wasteful aspect of the latency arms race is how such competition alters the public’s perception of the fairness of equity markets. Michael Lewis, Flash Boys author, made famous the notion that U.S. equity markets are “rigged,” when in fact

---

60 Chen, et al. (2016) study the effects of a speed bump introduced on the TSX Alpha system in Canada and show that the system decreases aggregate liquidity via its effect on other market centers.
many of the examples he cited were instances of market participants competing according to rules laid out by the SEC. Such confusion is far from harmless, as a loss of faith in the fairness of our capital markets can lead to decreased investment, higher cost of capital, and real effects on public firms. It can also give rise to intervention Congress in ways that are less than productive. If legislators discover their constituents believe they are not getting a fair shake in U.S. capital markets, history has shown they will not hesitate to act and unilaterally impose legislative solutions where such solutions may not be warranted.

Order types

One does not have to go too far back in time for the basic trading decision with respect to order type to be between a market and a limit order, with perhaps additional consideration given to a few rarely used modifiers such as immediate-or-cancel, all-or-none, or fill-or-kill. Today the situation is far more complex. Public exchanges typically have between 25 and 50 different order types, many of them either rarely used or used only by very sophisticated traders such as high frequency traders to accomplish targeted goals. According to KCG, when the number of exchanges is coupled with the number of order types offered by each exchange, a trader has a choice between over 300 order types and exchange destinations. KCG has created an Appendix to describe various order types that runs to 15 pages.

Why are they so many order types? There is likely more than one reason for this. One reason relates to the different demands of the various constituencies that trade in any market.

---

61 In a November 2012 Traders magazine article, Chris Concannon, CEO of BATS, is quoted as saying that BATS Global Markets itself as more than 2,000 order types. It is not clear how he arrives at this high a number, and it may be the result of computing the permutations that result from the various input parameters that define an order.

center. High-touch traders or traditional algorithmic traders may only make use of a very small number of order types, such as market, limit, ISO, and IOC. High frequency traders on the other hand are able to use the other order types tailored for particular situations that the HFT strategies are designed to address. Second, a number of these order types arise because of the complex working of Regulation NMS and its protected quotes. For example, orders linked to price sliding generally stem from Regulation NMS’s prohibition on displaying orders that lock or cross markets. The liquidity seeking order that wishes to avoid paying take fees would normally immediately execute if, when it was placed, it was immediately marketable. Price sliding algorithms reprice the order a minimum variation below the contra side of the market and continuously adjust the limit price so as to avoid crossing the market and paying take fees, while still preserving market display at the allowable minimum variation price. There are many variations on the price sliding algorithms.

There are two primary concerns with the large number of order types. First, they add substantial complexity to the marketplace, making it difficult to understand how all these orders interact. As a case in point is the BATS error in allowing its own price-sliding orders to trade-through the protected quotes in the current NBBO, in violation of Rule 611 of Regulation NMS. While the economic magnitude of this mistake was small, it is representative of the complexity in today’s marketplace. The second concern is the advantage that a large array of complex order types affords to sophisticated traders and institutions over less skilled market participants. The more sophisticated traders can make greater use of these various order types

---

and, in conjunction with their low latency trading systems, gain a meaningful advantage over less adept traders.

Note that in such a world it is not enough for a non-HFT market participant to merely content itself with a small set of order types. As an example, consider the case of a “hide not slide” order.\textsuperscript{64} It works similarly to the price sliding order described above with one important exception. Whereas the price sliding order will initially reprice the displayed portion of the order to be one minimum variation below the contra side of the NBBO, “hide not slide” orders will simply go dark and not display at all when the order locks or crosses the market. When the NBBO moves so that the original order no longer locks or crosses, it will not only be redisplayed but will move to the front of the queue at the display price, achieving an advantage over other orders that were active in the market and tracking the spread through price sliding over the ensuing period. A less sophisticated trader making use of simpler order types, including the simple price sliding order, would not be aware that they were sacrificing their queue positions to these “hide not slide” orders.

It also appears that some market center operators were less than clear with their market’s participants about available order types. For example, in 2015 the SEC charged UBS with failing to disclose to all subscribers of one of its ATSs the existence of an order type called the PrimaryPegPlus, which allowed certain subscribers to buy and sell securities by placing orders priced in increments of less than one penny. The SEC stated the UBS was prohibited from accepting such orders under Regulation NMS. By doing so, UBS allowed users of the PrimaryPegPlus order to jump ahead of other orders placed in conformance with Regulation

\textsuperscript{64} For a description of how this order type works, see https://www.wsj.com/articles/SB10000872396390444812704577605840263150860.
NMS rules at whole penny increments. In another example, Direct Edge Holdings, which operates the EDGEA and EDGEX exchange markets, settled charges with the SEC that they had failed to disclose the operation of price-sliding orders on their exchanges. The SEC charged that while exchange rules described a single price-sliding order, there were fact three different types of price-sliding orders in place including a Price Adjust, a Single Re-Price, and the Hide Not Slide order type. The exchanges were alleged to have selectively disclosed information about order types to a subset of members, creating a risk that not all markets in a participant would understand how exchange order types operated.

In June 2014, SEC Chairman Mary Jo White gave a speech entitled “Enhancing Our Market Equity Structure” in which she requested that the equity exchanges conduct a comprehensive review of their order types and how they operate in practice, and as part of this review, consider appropriate rule changes to help clarify the nature of their order types. Though many of these order types exist to facilitate compliance with the various strictures of Regulation NMS, it may also be possible to comprehensively reduce collective complexity of the array of order types. While some of these types may serve to reduce message traffic by, for example, automating cancel and replace functions induced by the order protection rule, others may be easier to eliminate. IEX, which only began operation to year ago, has on its rulebooks only five types of orders. Perhaps other exchanges and ATSs can be induced to follow suit.

III.D Best execution considerations

67 IEX order types are Market, Limit, Midpoint Peg, Primary Peg, and Discretionary Peg. See IEX Rule 11.190.
Brokers-dealer who handle customer orders have an obligation to obtain best execution for the orders. Though not precisely defined in any SEC rules, the obligation derives from a common law agency duty obligating an agent to act exclusively in the principal's best interest. This is true regardless of whether the broker-dealer trades the customer order in an agency or principal capacity. FINRA Rule 5310 on Best Execution provides that “[i]n any transaction for or with a customer or a customer of another broker-dealer, a member and persons associated with a member shall use reasonable diligence to ascertain the best market for the subject security and buy or sell in such market so that the resultant price to the customer is as favorable as possible under prevailing market conditions.”

While obtaining the best price for the customer is of paramount importance, the rule also contemplates consideration of other factors including the possibility of price dis-improvement, execution speed, likelihood of execution of limit orders, and customer needs and expectations. In addition, the SEC requires that “…broker-dealers deciding where to route or execute small customer orders in listed or OTC securities must carefully evaluate the extent to which this order flow would be afforded better terms if executed in a market or with a market maker offering price improvement opportunities. In conducting the requisite evaluation of its internal order handling procedures, a broker-dealer must regularly and rigorously examine execution quality likely to be obtained from the different markets or market makers trading a security. [Emphasis added]”

Thus, though the number and type of trading venues and market centers change, speed...
increases, and technology evolves, the obligation of a broker-dealer for best execution of customer orders remains a constant.

Certain institutional features of present-day U.S. equity markets pose challenges for the consistent application of best execution principles. Many of these features arise from legacy practices that began years or even decades ago when the structure of equity markets was quite different than it is today. Yet these institutional practices persist and are allowed under securities laws and current interpretations of best execution. We detail two of these institutional features below.

Payment for order flow

Payment for order flow is a long-standing practice whereby a market center offers to pay a routing broker compensation in exchange for sending orders to the venue. One of the earliest practitioners of payment for order flow was third-market firm Madoff Securities. When stocks traded in increments of 1/8th of a dollar, Madoff offered to pay firms with large amounts of customer orders, such Charles Schwab and Fidelity, $.01-.02 per share in exchange for sending their customer orders to Madoff. Firms such as Madoff were interested in purchasing such flow because retail customer order flow is uninformed, allowing market makers to avoid adverse selection and profitably trade at quotes determined in the exchange market. Though the practice raised concerns about conflicts of interest, it was allowed by the SEC because it reasoned that such payments didn’t necessarily violate best execution obligations since customers obtain prices at least as good as they would have obtained in exchange markets.

---

70 See Glosten and Milgrom (1985) or Easley and O’Hara (1987) for examples of the effects of informed trading on market prices and quotations.
Furthermore, if the market for order flow payments was competitive, customers of retail brokerages might expect to see the benefits of such payments in either in the form of lower trading commissions or improved services at the introducing brokers. The SEC required these payments to be disclosed to investors on confirmation statements, account statements, and new account opening forms.\footnote{Payment for Order Flow, Final Rule, (Release No. 34-34902; File No. S7-29-93), April 3, 1995.}

Today, some of the largest firms that “internalize” retail orders include Citadel Securities, KCG Holdings, and UBS Securities. These are sophisticated firms that internally trade against purchased customer orders, taking the opposite side of the customer were possible, and routing out to market centers in other instances. The firms use sophisticated computer algorithms and monitor multiple market centers in effort to discharge the duties of best execution that they inherit from the introducing brokers who route their order flow to them. The SEC has made it clear that such firms do not discharge their best execution duties merely by guaranteeing to fill customer orders at the NBBO. The internalizing broker-dealers must offer customers a chance of price improvement or execution at prices superior to the best bid and offer in the national market.

Of course, inferring what these prices are is a difficult exercise. For example, prices referenced in such calculations require determination of the NBBO, which comes from the SIP data feeds. This is assembled by these data processing utilities, who aggregate data from exchange venues, compute the best prices across all markets, and disseminate these data in a uniform nondiscriminatory manner. The SIP prices generally form the basis for determining the best prices in the market for most internalizers. However, proprietary firms that trade against...
incoming order flow are not limited to obtaining only SIP data—they are free to subscribe to private linkages from various market centers. SIP data contain only top of book prices and depth at each market center. Private linkage direct feeds, however, can contain detailed depth of book information and are generally disseminated with lower latency than the SIP data feeds.

Because the determination of what constitutes the best prices available in the market must be made on a continuous basis, the actual obligations of the broker-dealer with respect to best execution are not well-defined. The industry has generally interpreted the SIP NBBO to define the best prices available in the market, but it is not clear the regulator shares this view. Uncertainty around the benchmark for determining a broker-dealer’s best execution obligations is unproductive and should be clarified by the SEC or FINRA. Conflict around this situation was made clear in the recent administrative action against an internalizing broker-dealer. The SEC charged that the broker-dealer made misleading statements to its introducing broker-dealer clients about the manner by which it would seek to obtain the best price for customer orders, either through internalization or via routing to other market centers.72

The situation is further complicated because of how the market data are priced. The SIP data are priced through a process that is regulated by the SEC. For enhanced market data that contain depth of book and other information, however, the SEC has taken a lighter hand with respect to pricing and has generally left it to market forces to determine these prices. As such, OTC market makers and internalizers may not consume direct feed data from all market centers, leading different internalizers to compute different contemporaneous market prices.

While the SEC could act to prohibit the practice of payment for order flow, and the attendant practice of internalization, given that it has been allowed for over 20 years an outright

---

prohibition seems unlikely at this point. A possible solution is to more precisely define the obligations of the various parties to these transactions. First are the obligations of the introducing broker that receives the payment for order flow and routes customer orders to the internalizing dealer. Such brokers do not merely transfer their obligation of best execution to the executing broker--they retain responsibility for the original order. The terms of trade offered by the internalizing broker are usually spelled out in a precisely written contract between the two parties. Appropriate disclosures to end investors, including disclosures about the economics of the payment arrangements, seem a likely place to begin the improvement.

With respect to the executing broker, the situation is more difficult. These firms have large scale internalization execution processes. They are difficult for a regulator to audit, and thus assurances that a customer’s order received appropriate treatment are difficult to substantiate, and often come down to question of what an order would have received under some counterfactual execution practice. Ultimately the answer may lie in more refined disclosures of execution quality. At present, execution quality data in the form of Rule 605 reports for market centers, and Rule 606 reports on broker routing, are produced on a highly aggregated basis. Data are produced monthly and are binned relatively coarsely. If more detailed and precise execution quality reporting is produced, this may facilitate both better audits of executing brokers and enhance competition between internalizes. This was the original purpose of the predecessor execution quality reports, but given the changes in the technology of trading, the Rule 605 and 606 reports no longer effectively serve this purpose.

---

73 See Disclosure of Order Execution and Routing Practices, SEC Release No. 34-43590; File No. S7-16-00, creating the 11Ac1-5 and 11Ac1-6 (“Dash 5” and “Dash 6”) reports.
**Maker-taker fees**

Another form of payment for order flow that occurs on market centers such as exchanges and ATSs is a maker-taker fee. This fee model is a pricing structure whereby market centers charge traders a “take” fee for removing liquidity by trading against a resting order, and rebate to traders a “make” fee for placing liquidity on the market’s limit order book that subsequently gets executed.\(^7^4\) Rule 610 of Regulation NMS requires these fees to be no more than 0.3 cents per-share. However, maker-taker pricing is not a consequence of Regulation NMS as the practice predated the rule by at least a decade when the Island ECN used a maker-taker pricing model to attract order flow to its book.

Maker-taker pricing in today’s market is a dynamic process, and traders with order flow to route pay careful attention to the pricing charges and rebates. Maker-taker pricing is of concern for a number of reasons, and a complete treatment of the subject is far beyond the space pages that can be allotted to the topic in this paper. However, one of the most important concerns is that these payments create a conflict for brokers who have an obligation to seek best execution for their customers’ orders. Instead of routing orders to the markets with the best expected outcome, brokers might seek to maximize the value of the rebates they receive, while minimizing the cost of the fees they pay. This is true for both marketable and nonmarketable orders. Battalio, et al., (2015) examined the routing decisions for nonmarketable orders by a group of brokers and found that a number the brokers routed these nonmarketable orders to the market centers that paid the highest rebates. The authors felt that the data indicated that these routing decisions were inconsistent with the brokers’ duty of best execution. At a Senate hearing

---

\(^7^4\) There are also venues that use an inverted fee schedule, charging for “making” liquidity and paying rebates for “taking” liquidity.
in 2014, a senior executive of one of the routing firms, TD Ameritrade, testified that the firm had a policy of sending its nonmarketable orders to market centers that paid the highest rebate. While there has been some industry critique of the Battalio study, it seems clear that the potential for such a conflict is very real.

Proponents of maker-taker pricing argued that it is a competitive reality of today’s marketplace and that exchanges compete vigorously along these dimensions. Were maker-taker pricing to be prohibited for public exchanges, they would be at a disadvantage relative to unregistered (as an exchange) venues such as broker-dealers and ATSs that do not have prohibitions against paying inducements or charging fees. If maker-taker payments were prohibited, volume might migrate to internalizers and dark ATS venues, which, as discussed earlier in the section, would be concerning to regulators and other market participants.

Finally, critics have observed that because of Regulation NMS prohibitions against displaying sub-penny quotations, maker-taker fees are not reflected in market quotations, harming pricing transparency and violating the spirit of the firm quote rule. The price you see at a given market may not be the net price you will pay if you trade that market. They argue that either prices should be inclusive of these fees, if they are to be allowed at all.

If the SEC decides that maker-taker pricing is a problem they needs to resolve, there are another number of paths that could be taken. First, there could be a requirement that fees paid and rebates received be passed through to the ultimate investor. Second, the 0.3 cents per share limit could be reduced or set to zero, eliminating maker-taker pricing. Finally, the requirements against quoting sub-penny increments could be relaxed and the maker-taker fees could be reflected in displayed quotes. These solutions have been previously discussed and have been found wanting for various reasons.
More realistic perhaps is to approach the problem from the perspective of disclosure. Executing brokers, and the introducing brokers that route to them, could be required to disclose their routing policies with respect to maker-taker fees. Executing brokers would be required to disclose how they factor the existence and size of rebates and fees into their routing decisions. Because this routing is automated and is encoded into software, the disclosure should be readily verifiable. The introducing brokers would be required to disclose how the maker-taker routing practices of the various potential executing brokers affect their choice of executing broker. This disclosure could be further enhanced by quantification of the per-share rebates earned and taker fees paid by the executing broker. In addition, introducing brokers could be required to disclose how their policy toward executing broker selection is affected by payment for order flow fees received.

Regardless of which solution, if any, is taken, brokers retain the responsibility for best execution toward customer orders. FINRA and the SEC could consider issuing additional guidance with respect to best execution obligations as they relate to maker-taker fees. It is not clear what form such guidance would take, but such disclosure would allow exchanges to retain the competitive benefits of maker-taker pricing vis-à-vis the less regulated ATS and dealer markets centers.

* * * * * * *

As the above discussion shows, there are number of challenges concerning the application of the duty of best execution to today’s markets. Yet the very flexibility of the agency duty may in fact be its most valuable trait. The obligation to seek the best terms of trade for the order is, by its very definition, flexible and adaptable to the current market environment. Brokers, and especially brokers who route high volumes of retail orders, are the market
participants most able to make judgments about which venues are best for customer orders at any one moment in time. Rather than viewing best execution as a problem for today’s environment that needs to be solved, it may be better to view it as the solution to the thornier market structure problems. A renewed and vigorous approach on the part of the SEC and FINRA to best execution oversight may cause introducing and executing brokers to be increasingly vigilant of their routing decisions for retail orders. Regulators could make it clear that they expect such brokers to be fully informed on issues such as current maker-taker pricing structures, order forms, adverse selection intensity, and other features that determine how well a given retail order will fare in at any particular market center.

Such an approach may require reconsideration of a basic aspect of the application of best execution principles. To date regulators have not generally looked at the best execution on an order-by-order basis, but have simply required that brokers periodically “regularly and rigorously” assess their routing decisions, considering both venues to which orders were routed and those to which they did not route. Given the advances in routing technology and the fact that routing decisions are now made on an order-by-order basis, perhaps best execution requirement should also begin to reflect this order-by-order feature. FINRA has already contemplated such standards for larger-sized orders, but has yet to bring such standards to smaller retail orders.75 Were such a proposal to be made by regulators, it should surely raise substantial concern among industry participants. Great care would need to be taken to avoid creating an environment where executing brokers are judged on an ex-post basis and held responsible for statistically unfavorable outcomes for retail orders.

75 See NASD Notice to Members, 01-22, Best Execution, April 2001, at footnote 13. Also, see discussion in FINRA Regulatory Notice 15-46, Best Execution, November 2105, pg. 3.
III.E System robustness

The success of our equity markets depends on a belief that trading venues will operate in a reliable and fair manner. In recent years, however, there have been numerous widely reported events that have raised concerns, such as: the May 6, 2010 flash crash; system shutdowns at NYSE, NASDAQ, BATS and other venues; and fines levied against high frequency traders, trading venues, and other market participants for improper behavior. As well, technological advances have increased the likelihood that a seemingly minor operational problem at a single entity can spread rapidly across the entire system and cause harm to a wide range of market participants.

These concerns should be taken seriously. In today’s highly charged, populist political environment, the public’s perception of market robustness and fairness may matter more than the reality. If the tide of public opinion decides that the market is rigged or unfair, then there is the risk that market rules and regulations will be set based on political calculations, rather than based on objective evidence and careful deliberations.

To address concerns about system disruptions and other technological failures, in 2014 the SEC adopted the Regulation Systems Compliance and Integrity (Regulation SCI). This regulation is designed to: (i) reduce the occurrence of systems issues; (ii) improve resiliency when systems problems do occur; and (iii) enhance the SEC’s oversight and enforcement of securities market technology infrastructure.76 A large component of Regulation SCI is requiring certain entities to have written policies and procedures to deal with issues related to the capacity,

---

integrity, resiliency, availability, and security of their systems.\textsuperscript{77} In addition to the prevention of system issues, Regulation SCI considers how to contain and minimize problems should they occur.

Regulation SCI applies to “SCI entities” which include SROs, alternative trading systems with trading volumes exceeding specified volume thresholds, disseminators of consolidated market data, and certain exempt clearing agencies. Since each system disruption has unique causes, it is a challenge for the regulation to define what constitutes a “major SCI event” requiring the SCI entity to disseminate information about the disruption. Risk management after an event can be slow relative to the speed of the overall market.

Importantly, Regulation SCI would not apply to HFT market makers and broker-dealers. As such, had the rule been in effect in 2012, it would not have protected the market from the severe adverse effects of the accidental inclusion of obsolete code in the automated routing system at Knight Capital. In that incident, while handling just 212 small retail orders, the system routed millions of orders into the market over a 45-minute period, obtaining 4 million executions in 154 stocks for more than 397 million shares.\textsuperscript{78} The resulting unwanted long and short positions in these shares led to over $460 million in losses for Knight Capital, and perhaps more importantly, caused severe price dislocations in many securities.\textsuperscript{79}

The scope of Regulation SCI is limited and does not cover many of the most pressing concerns about system robustness and integrity in U.S. trading markets. In this section, we consider five additional areas of potential concern: (a) flash events; (b) episodic liquidity and the

\textsuperscript{77} See page 2 of SEC Release No. 34-73639.


\textsuperscript{79} For 75 of the stocks, Knight’s executions comprised more than 20 percent of the trading volume and contributed to price moves of greater than five percent. (Source: p. 6, SEC Administrative Proceeding File No. 3-15570, “In the Matter of Knight Capital Americas LLC Respondent.” https://www.sec.gov/litigation/admin/2013/34-70694.pdf)
role of designated market makers; (c) potential breakdowns in the creation and destruction mechanism for exchange traded funds; (d) quote stuffing and latency arbitrage; and (e) cybersecurity. We conclude by examining the potential implications of using blockchain technology in trading markets.

Flash events

Market breaks are not a new phenomenon. An entire chapter (Chapter XIII) of the original Special Study focused on the market break of May 28, 1962. As these two passages from the study illustrate, many of the issues in 1962 are still relevant today:

The avalanche of orders which came into the market during this period subjected the market mechanisms to extraordinary strain, and in many respects they did not function in a normal way. Particularly significant were the lateness of the tape and the consequent inability of investors to predict accurately the prices at which market orders would be executed. (p. 859)

[...]

The history of the May 28 market break reveals that a complex interaction of causes and effects—including rational and emotional motivations as well as a variety of mechanisms and pressures—may suddenly create a downward spiral of great velocity and force. (p. 861)\textsuperscript{80}

In today’s markets, one area of concern has been whether high frequency trading promotes sudden and unexpected price dislocations, which are popularly referred to as “flash crashes.” Public attention first focused on flash crashes with the events of May 6, 2010, when an algorithm rapidly sold 75,000 S&P500 e-mini futures contract. The index was already down 4% by the time the large sell order hit at 2:30 PM. The index fell a further 5 to 6% and then recovered most

\textsuperscript{80} http://3197d6d14b5f19f2f440-5e13d29c4016e96ebeb0197c579b45.r81.cf1.rackcdn.com/collection/papers/1960/1963_SSMkt_Chapter_13_1.pdf
of the decline within the space of about a half hour. In the initial joint report of the CFTC and SEC, these price movements were attributed to a sequence of events including the exhaustion of the liquidity supply high frequency traders, traditional buyers, and cross market arbitrageurs who spread the price pressure to other markets. Eventually a “hot potato” effect developed where blocks of futures contracts rapidly moved among the same set of traders. Depth fell, liquidity vanished, and prices crashed. When a five second pause was triggered on the CME, prices began to recover and within minutes they had risen to almost their previous levels. In a follow-on report several months later, the SEC made a set of recommendations including the implementation and coordination of “limit up/limit down” trading halts, pre-trade risk safeguards to prevent access the markets without appropriate risk controls, and improved controls over algorithmic trading strategies.

On October 15, 2014, the U.S. Treasury market severely dislocated in a “melt-up” during which over a 13-minute period the yield of the 10-year government bond experienced a 16 basis point drop and subsequent rebound. According to a subsequent joint regulatory report, HFT firms withdrew from their market-making function as the dislocation proceeded, causing liquidity to fall. This in turn caused different business units of individual high frequency trading firms to unintentionally engage in self-trading. This higher trading volume slowed the pace of trading. The report cited no single factor is the cause of the event.

82 SEC Release 34-67091.
83 The SEC also alleged in a civil complaint that spoofing activity of a U.K based trader caused artificial prices to exist (CFTC v. Sarao (2015))(CFTC v. Sarao, 2015a, para. 2) as quoted in Aldrich]
On August 24, 2015, broad-based exchange traded funds (ETFs) declined sharply before and at the 9:30am opening of the market. The $65 billion iShares Core S&P 500 ETF fell by over 25% in the opening minutes of the session,\(^85\) and the equity of KKR fell by almost 60% from its previous close before recovering. Though the exact causes of this dislocation are still be debated, the operation of “limit up/limit down” halts put in place after the May 2010 flash crash appear to have played a role.\(^86,87\)

Here, too, there is an academic literature on the subject, but unlike the topics of liquidity and pricing efficiency, the evidence is somewhat more mixed. Kirilenko, Kyle, Samadi, and Tuzun (2017) show that the trading patterns of high frequency traders did not change as prices fell during the May 6, 2010 event. Aldrich, Grundfest, and Laughlin (2017) conclude that instability of the market data infrastructure contributed to the May 2010 flash crash, and emphasize that in a market dominated by algorithmic and high frequency trading, data integrity becomes a paramount issue. Menkveld and Yueshen (2016) attribute the cause on the 2010 flash crash to the breakdown of cross-market arbitrage mechanisms. Because the seller was forced to find sellers only in the E-mini market, this concentrated the effect of the decline. They point out that failed intermarket linkages can be costly for investors in highly fragmented markets. Easley, López de Prado, and O’Hara (2012) attribute the 2010 flash crash to the combination of automated market makers and increased order flow toxicity, which combined to cause market

\(^85\) We consider the role of ETFs in more detail in a later section.
\(^86\) The October 7, 2016 UK Sterling flash event is another example of a sudden market dislocation. For details, see BIS (2017).
\(^87\) Trading halts may be a problematic solution to flash events. As an alternative, Bethel, et al. (2011) explore the feasibility of an early warning system to predict flash crashes. When a flash crash is likely, a mechanism could be used to “slow” the market down, rather than using an abrupt trading halt.
makers to withdraw their quotes and liquidate positions.88 Interestingly, the authors propose contracting around liquidity provision conditional on times of high adverse selection.

A notable contrast to the above papers is the study by Golub, Keane, and Poon (2012). The authors study four months of volatile prices over a period of six years and find that there were over 4,500 mini flash crashes in individual stocks over this period. The attribute the cause of these mini crashes to use of intermarket sweep orders that scrape liquidity from the top-of-book market centers. Once these protected quotes are taken out, NMS rules permit traders to execute against the remaining unprotected orders of individual market centers, eating into the limit order book and ignoring superior prices available on other exchanges. They attribute these mini crashes to HFT and fragmented markets.

The lack of clear consensus by both regulators and academics as to the root causes of the various flash crashes suggests more work needs to be done. This should not be surprising given the tightly-coupled nature of the systems depend so integrally on data feeds, order types, and the pre-programmed behavior of automated trading systems. A potentially complicating aspect of flash crashes is the ability of SROs and regulators to break previously agreed-upon trades based on the notion that these trades were done at “clearly erroneous” prices.89 If high frequency traders and their algorithms believe they have entered a period in which there is an increased probability that consummated transactions will, on and ex-post basis be negated on the basis of such a regulatory finding, they will not be able to appropriately manage the risk of their trading

88 Easley, Lopez de Prado, and O’Hara (2012) develop a measure of the toxicity of order flow, which captures the extent to which (algorithmic) market makers may be unknowingly providing liquidity at a loss. The measure is calculated by grouping consecutive trades across time according to the time required to trade an exogenous level of order flow. Within each group, the sign of price changes over one-minute periods is used to infer buyer- versus seller-oriented trades, which is then used to create a high frequency analog of the popular PIN measure. Easley, et al. show that this measure (VPIN) is a useful indicator of short-term volatility.
89 For example, see NASDAQ Rule 11890. Jurich, Maslar, and Roseman (2017) examine the effect of uncertainty regarding the possible cancellation of erroneous executions in U.S. markets.
operations and hedge their positions. They can logically be expected to pull back from their market-making function, exacerbating liquidity issues at a time when market depth is most critically needed. That the infrequency of the major flash crashes makes their study more challenging should not dissuade regulators and researchers from seeking to better understand these complex interactions.

Episodic Liquidity and Designated Market Makers

Flash crashes, both market wide and stock specific, are a dramatic manifestation of episodic liquidity. But even in the absence of flash crashes, liquidity can be fleeting in an environment in which its provision is primarily via high frequency traders, who can enter and exit the market as market conditions change. For this reason, it is important that market quality measures are considered in the time series, not only in averages.

The measurement of liquidity itself can be challenging in modern markets. Quote-based measures of liquidity, such as bid-ask spreads, effective spreads, inside depth, and round-trip costs, are less relevant to institutional traders in a world with fleeting quotes that have tiny spreads and little depth. To address this, Barardehi, Bernhardt, and Davies (2016) develop a new measure of liquidity based on the average per-dollar price impact of trading a fixed-dollar position of institutional trading size. Using this measure in an asset pricing framework, they show that liquidity premia have risen somewhat in recent years, suggesting that institutional traders are pricing in higher liquidity risk.

Prior to the changes triggered by Regulation NMS, liquidity and price stability were provided by NYSE specialists and NASDAQ dealers. Those roles have been supplanted by high

---

90 Carlin, et al. (2007) provide a theoretic model of episodic liquidity.
frequency traders acting as market makers. Most of these market makers have no formal obligations or designations, although the NYSE has designated certain Supplemental Liquidity Providers (SLPs) that are given monetary incentives to provide liquidity in assigned securities.\footnote{https://www.nyse.com/publicdocs/nyse/listing/fact_sheet_slps.pdf} These market makers, however, do not have an affirmative obligation to provide liquidity and as a consequence, they may stop providing liquidity under certain market conditions. In such an environment, it is worthwhile asking whether human market makers can still play a role.

On the NYSE, the former specialist role has been effectively replaced by designated market makers (DMMs) that have mild obligations to maintain a fair and orderly market.\footnote{https://www.nyse.com/publicdocs/nyse/listing/fact_sheet_dmm.pdf} Their obligations are so loosely defined that the obligations unlikely to have substantial impact. Clark-Joseph, Ye, and Zi (2016), however, use an exogenous shock to the NYSE system to show that even with relatively weak obligations, these DMMs improve the overall market quality. Clark-Joseph et al. do not have a conclusive explanation for why the DMMs provide liquidity under such weak obligations, but conjecture that the DMMs compete with each other on reputation for preferred stock allocations.

DMMs appear to play a valuable role in Canada as well. Anand and Venkataraman (2016) compare liquidity provision by endogenous liquidity providers and designated market makers on the Toronto Stock Exchange, and show that the DMMs reduce execution uncertainty by participating in undesirable trades.

Some European markets have DMMs that are required, by contract, to keep the spread within certain binding limits. These DMMs receive compensation for providing this service. Anand et al. (2009) finds that, on balance, firms contracting with designated liquidity providers
on the Stockholm Stock Exchange experience a decreased cost of capital and significant improvements in market quality and price discovery.

Until recently, issuers in U.S. markets have not been able to enter into a contract arrangement with a DMM to support the liquidity of their stock. In 2013, the SEC approved programs on a pilot basis to allow issuers of exchange-traded products to compensate market makers in those securities. It is unclear whether these programs have been successful, as one of the largest ETF providers, BlackRock, announced in July 2016 that it was withdrawing from the program. In light of this decision and other evidence on episodic liquidity, regulators and trading venues must decide what role designated market makers, with an affirmative obligation to provide liquidity, have in modern financial markets.

**Exchange Traded Funds**

Exchange Traded Funds (ETFs), and more generally, Exchange Traded Products (ETPs) are enormously popular. ETFs now have more than $2.614 trillion in assets. Credit Suisse estimates that ETFs account for 30% of all U.S. trading by value, and 23% by share volume. More and more retail investors are electing to buy ETFs rather than have direct ownership of individual stocks or owning open-end mutual funds. Notably, the vast majority of ETFs, by assets under management, are issued by only four firms (Blackrock, Vanguard, State Street Global Advisors, and Invesco Powershares). The magnitude of trading in ETFs, combined

---

93 Dolgopolov (2013) provides an overview of the regulatory framework prohibiting issuer-to-market maker compensation, and the motivation for the pilot ETF programs.
96 ETFs are eating the US stock market. Financial Times. January 24, 2017. https://www.ft.com/content/6dabad28-e19c-11e6-9645-c9357a75844a
with their special features, have led to them to be a potential source of stress for overall market conditions.

Authorized participants ensure that the price of an ETF tracks that of its constituents through a complex process of creation and redemption. This arbitrage process can breakdown when there are trading halts or other dislocations in the market. The magnitude of ETFs combined with breakdowns in the arbitrage relationship have introduced an important potential source of fragility in trading markets. Ben-David, et al. (2016) find that during turbulent market episodes, arbitrage is limited and ETF prices diverge from those of the underlying securities.98

On August 24, 2015, delays in the opening of trading for many stocks, combined with a trading pause in S&P 500 Index futures, lead to a breakdown of the arbitrage relationship between ETFs and their constituents. Over the next 30 minutes, there was a crash in the prices of many ETFs, resulting in prices that were far below those of their no arbitrage values. Limit Up-Limit Down (LULD) trading halts appear to have compounded the problem. 327 ETFs were hit with five-minute trading halts that morning, with 11 ETFs halted 10 or more times.99 Notably, many of the LULD halts occurred in smaller ETFs with lower turnovers. Gerig and Murphy (2016) find that ETFs in the bottom quartile of turnover were three time more likely to pause than those in the top quartile. Furthermore, Gerig and Murphy find that ETFs with high correlations with the S&P 500 index were much more likely to pause, and that ETFs within the top quartile of S&P 500 correlations 21 times more likely to pause, than those in the bottom

---

98 In the context of bond ETFs, Pan and Zheng (2017) show that the arbitrage relationship can breakdown further when there is a liquidity mismatch between the ETF and the underlying bonds, and the ETF’s authorized participants have the incentive to use ETF creations and redemptions to manage their own bond inventory imbalances.

quartile. These ETFs with high S&P 500 correlations experienced both large volume spikes and large liquidity drops.

The events of August 24th suggest that regulators need to think carefully about the impact of trading halts on ETFs, and about the breadth and diversity of holdings of an ETF. Of the ETFs focused on domestic equity, the ICI classifies 397 as broad-based and 388 as sector/industry based.\textsuperscript{100} Some of the sector/industry ETFs are highly specialized, with a small number of constituent holdings. For a narrow ETF, a breakdown of the creation/destruction arbitrage relationship is more likely to occur when some of the underlying securities are halted.

ETFs raise other issues that impact market integrity and performance. For instance, narrowly defined ETFs can be used to engage in manipulation and insider trading, in a practice known as ETF-stripping, by which a trader buys the ETF that includes the stock of interest and then short sells all of the other stocks in the ETF, or vice versa.\textsuperscript{101} Also, ETFs tend to increase stock return co-movement (see Glosten, et al. (2016)), which reduces the ability of liquidity providers to diversity their stock specific risks, and thereby increases volatility across stocks.

**Quote Stuffing**

Quote stuffing refers to a high frequency trading strategy in which a very large number of orders to buy or sell securities are placed in quick succession and then canceled almost immediately. As described in Credit Suisse (2012a,b), quote stuffing can be done for three reasons: (i) it can be used to game other orders, such as peg-to-mid orders, that base their pricing on the best bid and ask prices, and subsequently take advantage of these orders; (ii) it can be

\textsuperscript{100} Source: Investment Company Institute. https://www.ici.org/research/stats/etf/etfs_01_17 (January 2017)

used to create false mid-points near the prior bid or ask, which are then used to trade in a dark pool that uses the mid-point as its reference price; (iii) it can be used to create stale prices and slow market data, which allow the quote stuffer to take advantage of other market participants’ slower connections and create opportunities for latency arbitrage. Latency arbitrage refers to a practice whereby high frequency traders profit by simultaneously buying at the ask on one market and sell at the bid on the other market during short instances when the best quotes on those two venues are crossed. While quote stuffing could be a form of manipulation, it is also important to note that it could be a natural by-product of two algorithms interacting with each other but failing to converge to a stable equilibrium. Regardless of the cause, quote stuffing impacts market integrity by clogging message traffic and preventing other traders from updating or submitting their orders.

Quote stuffing is challenging to identify empirically, since it is unclear whether episodic spikes in quoting activity are the consequence of manipulation or a natural response to higher volatility. Gai, Yao, and Ye (2013) develop a clever identification strategy based on the fact that NASDAQ stocks are randomly grouped into six identical but independent data feed channels. They find that stocks in the same channel have an abnormal correlation in message flow, which is consistent with quote stuffing. Egginton, Van Ness, and Van Ness (2016) find that over 74% of U.S. exchange-listed securities experienced at least one episode of quote stuffing during 2010. They also show that such episodes have a negative impact on market quality, with targeted stocks suffering decreased liquidity, higher trading costs, and increased short-term volatility.

Despite evidence that quote stuffing is a relatively common occurrence, it appears that only one broker-dealer has been fined to date.\textsuperscript{102} Quote stuffing is related to other forms of

\textsuperscript{102}http://www.nasdaqtrader.com/content/marketregulation/NASDAQ/DisciplinaryActions/CDRG_NQ_2014.pdf
market manipulation often associated with HFT, such as layering, spoofing, price and venue fade, and momentum ignition. Regulatory proposals designed to limit quote stuffing and other related forms of manipulation include: (a) fines for exceeding a specified order-to-trade ratio; (b) fees to update quotes or a limit on the number of quote updates within a certain time period; and (c) a minimum resting time for orders.\(^{103}\) All of these proposals come with the risk that they will harm long-term investors and have other unintended consequences. More research is needed to establish whether the apparent costs of quote stuffing are sufficient to warrant these measures.

**Cybersecurity**

The frequency and sophistication of cybersecurity attacks is a major concern for trading venues and broker-dealers. Fortunately, major U.S. trading systems do not appear to have been directly compromised thus far. There have been some concerns, such as when NASDAQ announced in February 2011 that suspicious files had been found on its servers, although the files appeared to have only affected its web applications and did not compromise its trading systems.\(^{104}\)

In a recent FINRA survey on cybersecurity\(^{105}\), the top threat according to respondents was “Cyber risk of hackers penetrating systems for the purpose of account manipulation, defacement or data destruction, for example.” A hacker could shut down an exchange, tamper with critical market data, or use a financial intermediary to execute unauthorized trades. Direct market access by non-brokers, without proper risk management controls, may increase the

\(^{103}\) See Credit Suisse (2012a) for additional details.


The large number of trading venues adds to the cost and complexity of managing security and responding to a cyber-attack. At the same time, the existence of multiple trading venues provides a useful redundancy; if one trading venue is unable to operate due to a cyber-attack, malfunction, natural disaster, or terrorist event, another trading venue can serve in its place. Regulation SCI recognizes the issues with interconnected systems, and requires that written notice by provided to the SEC within 24 hours of any responsible personnel becoming aware of a systems intrusion. As well, SCI firms are required to develop a coordinated response to a breach, in which a process is established for the information security, technology, legal and compliance teams.

Much of the issues with cybersecurity come from the enormous complexity of the financial trading systems. Kirilenko and Lo (2013) argue that financial regulation should be designed with this complexity in mind, and adhere to four basic design principles: (i) regulation should promote best practices in systems design and complexity management, viewing automated markets as complex systems composed of software applications, hardware devices, and human personnel; (ii) effective risk safeguards should be consistent with machine-readable communication protocols, in addition to human oversight; (iii) financial regulation should make the design and operation of financial products and services more transparent and accessible to

---

automated audits; and (iv) financial regulation should encourage innovation to be platform-neutral, helping to avoid locking-in old technology and practices.

The risks of a cyber-attack are more than that of bringing down the trading system for a few days; there is also the long-term risks of corrupt records in the system. Distributed ledger technology, discussed below, appears to show great promise as a means to provide redundancy of trading records and thereby reduce the potential negative consequences of cyber-attacks.

**Blockchain**

Distributed ledger technology, commonly referred to as blockchain, has the potential to transform the payment, clearing, and settlement processes for securities trading.\(^{107}\) It can enable direct peer-to-peer trading, without the need for clearing intermediaries, and it can allow for settlement on a nearly instantaneous basis. The management consulting firm Oliver Wyman estimates revenue from clearing and securities services (settlement, custody, collateral management) alone to be $45 - $55 billion per year, so the potential cost savings from using blockchain technology to reduce market frictions are substantial.\(^{108}\)

Blockchain is best known as the technology behind Bitcoin, a virtual currency.\(^{109}\) A blockchain is a distributed ledger or database that contains a list of ordered records, called blocks. Each block contains a record of a transaction and a link to the previous block. The entire database is stored across all of the computers (or nodes) in a peer-to-peer network. The distributed network helps protect the system against hacking, since changes to the blockchain

---

\(^{107}\) Lee (2016) provide an overview of the potential of blockchain technology in securities trading.


\(^{109}\) The Federal Reserve is studying distributed ledger technology applications in payment, clearing, and settlement processes (see Mills et al. (2016)).
must be verified using a decentralized protocol based on the entire network of nodes. By design, blockchain has a built-in redundancy, and blocks cannot be modified once they are created and verified.

Because blockchain transactions occur in a peer-to-peer fashion, every transaction requires signature verification to prove its source. From a security perspective, this process is a potential weak link in the blockchain since there needs to be a consensus mechanism for verifying these signatures. Policies need to be established to handle these “virtual identities” and how much information they should reveal about the actual participants in the transaction.

Applications of blockchain technology in securities trading are already starting to emerge. For instance, blockchain technology was used for the recent preferred stock issuance by Overstock.com.110 DTCC will use blockchain technology to process credit default swaps.111 Nasdaq Linq uses blockchain technology to allow private-market securities issuers to complete and record private securities transactions.112 The Australian Stock Exchange has announced plans to use blockchain technology in a new post-trade system to replace the existing CHESS system for clearing, settlement and asset registration for equity cash markets.113 Lykke has developed a peer-to-peer foreign exchange trading platform based on blockchain.114 The Japan Exchange Group has established an internal working group to explore the pros and cons of using blockchain technology.115 A fintech company, R3, is leading a consortium of 80 of the world’s

114 https://www.lykke.com/technology
largest financial institutions to develop a new distributed ledger platform, Corda, for the banking industry, designed to provide privacy and limit duplication of records. Another platform, known as Hyperledger, is being led by the Linux Foundation.

A core feature of the traditional blockchain is that it is based on an open, permissionless network. In a public blockchain, there is a distributed network, with open access and irreversible blocks. The entire ledger is public, so while a user’s public key (or identifier) may not be easily connected to their real-world identity, it may be possible to identify patterns in the usage of a given public key in the transaction history which might allow one to infer information about ownership. The distributed network makes a public blockchain more secure since the information is stored across a potentially unlimited number of nodes and, in theory, an attacker would need to gain control of more than half of the network nodes to corrupt the ledger. An alternative form of blockchain is based on a closed, private network in which verification is authorized or permissioned by a central authority. In a private blockchain, the network has a select number of nodes, with filtered access and blocks that can be changed. In a private blockchain, ownership information can be public or private. A closed network is less secure, but there is more control on entrants. Importantly, signature verification in an open network can be a time-consuming process (at least in the context of high frequency trading), whereas in a closed network signature verification may not be necessary.

In the context of trading markets, Malinova and Park (2016) develop a model that illustrates the trade-offs between open and private blockchains. In their model, the key issue is whether the public identifier allows other traders to know the identity and size of their

\[\text{http://www.r3.com/}\]
\[\text{https://www.hyperledger.org/about}\]
\[\text{A market participant could reduce this risk by using a new public key after just a few transactions.}\]
counterparty. As is true in trading markets today, in some situations too much transparency can be detrimental to liquidity and the execution of large institutional sized positions, as other investors will trade ahead of this information. But despite this consideration, Malinova and Park (2016) show that the highest welfare occurs in environments with full transparency, in part because it enables large investors to contact other large investors directly.

Blockchain and related technologies are still in their infancy. In its current form, the decentralized nature of distributed ledger technology does not appear to be well-suited for aggregating orders in real-time due to the massive amount of information generated by the high speed of equity trading. As such, some market participants are already concerned about potential costs from having only some securities and only some parts of the trade lifecycle on blockchain. In particular, the payment system can prove to be a bottleneck in settlement process, unless a blockchain-based currency such as Bitcoin is also used.

Despite these concerns, there is genuine excitement about the potential uses of blockchain technology across diverse areas of finance and commerce. At the 2017 World Economic Forum Annual Meeting in Davos, there were a dozen formal and informal sessions on blockchain. \(^\text{119}\) Blockchain technology has the potential to transform the organization and regulation of trading by enabling direct “end-investor” to “end-investor” transactions. It is worth noting that one of the attractions of Bitcoin was the perception that its decentralized nature made the currency beyond the control of government. In this light, regulators may want to be proactive in establishing a framework for the control and access of information on distributed ledgers used

for trading securities. To this end, the SEC has already formed a working group to build expertise, identify risk areas, and coordinate efforts on distributed ledger technology.120

III.F Market data environment

Market data are the trade and quotation information associated with trade in equities.121 Core trade data consist of a time-stamped real-time record of the price and number of shares associated with each stock trade (i.e., the “ticker tape.”) Core quotation data consist of the top-of-book prices and depth of each registered public exchange, as well as the NBBO. These data are provided pursuant to several SEC-approved market data plans, are which are administered by a securities information processor (SIP). The exchanges provide the core data to the SIPS, who then compute an NBBO and distribute the consolidated data to the public. Currently NYSE Euronext and Nasdaq OMX Group each run a SIP on behalf of the public exchanges.

Trade in securities both produces and consumes market data. The data are produced when orders are placed on the exchange, which produces quotation information, and when trades occur on the exchange, which produces last sale information. Market data are consumed during the process broker-dealers follow when accepting orders from customers. A broker-dealer wishing to trade would need to know, at a minimum, what the bids and offers are at the various exchanges. In crafting its order placement strategy, the broker may very well also like to know the prices and quantities of recent trades. In addition, Exchange Act Rule 11Ac1-2, also known as the “Display Rule,” requires that any broker-dealer using quotation data be provided data that

---

120 At this stage, it is unclear whether some existing SEC regulations, such as those for transfer agents and clearing agencies, require registration for blockchain applications.

at a minimum contains either (a) the NBBO for the stock, or (b) a quotation montage for the stock from all reporting market centers. Similarly, a broker-dealer using last sale data must at a minimum be provided the price and volume of the most recent transaction in that stock from any reporting market center, as well as an identification of that market center. This means is that if a broker-dealer is in the business of executing orders for customers, they much must purchase and use the consolidated SIP data in transacting their business.

The effect of the above is to create an inelastic demand for core market data on the part of broker-dealers and other market participants. To ensure that such data are reasonably priced, the collection and pricing of these data must occur in a manner that is consistent with the requirements of Exchange Act and its various rules, following a process overseen by the SEC. This process insures that core data prices are reasonable. For example, for Network A (NYSE) securities, current monthly charges for core market data include price points at $0.0075 per quote, $1.00 for a non-professional user, and $1,250 for professional electronic access to quotations.122 The SEC has stated that fees for core market data “need to be tied to some type of cost-based standard in order to preclude excessive profits if fees are too high or underfunding or subsidization if fees are too low.”123

In addition to core data, market centers who route and execute orders produced a host of other enhanced market data products. These market centers recognize that market data is a key input for their trading constituency. The sale of data has become an important revenue source for the exchanges. For example, for BATS Global Markets, while transaction revenue net of

---

liquidity payments was $164.4 million, market data fees were $99.4 million.\textsuperscript{124} Nasdaq OMX data products revenue was $399 million for 2015.\textsuperscript{125}

Enhanced market data contain detailed information about the state of the market’s order book, including the price and share quantity of every displayed order on the book, broadcast in a variety of formats. These data are available at various degrees of latency, with higher cost associated with the lower latency products. For example, the NYSE Integrated Feed provides a microsecond precision order-by-order view of events in the NYSE equities market. Its pricing for non-display use (algorithmic and high frequency trading) can run to $20,000 per month.\textsuperscript{126}

There is no SEC or FINRA imposed regulatory requirement that broker-dealers accepting customer orders obtain or use such non-core data. Accordingly, the SEC has taken a much more hands-off approach to oversight of the pricing of non-core data products. This pricing approach has not been without controversy, however. In 2009, the SEC was sued by NetCoalition, a public policy corporation representing approximately 20 internet companies (including Google and Yahoo!) and SIFMA (a trade association representing more than 600 securities firms and banks) over the approval of NYSE ARCA data fees that NetCoalition deemed excessive.\textsuperscript{127}

While the environment around market data may seem quiescent, this is not the case. Though a complete treatment of market data questions is beyond the scope of this paper, we highlight below several key issues that we feel should be examined in any comprehensive reform of U.S. equity market structure.

\textsuperscript{124} BATS Global Markets S-1 filing, December 2015. 
\textsuperscript{125} Nasdaq Inc. Form 10-K, year ending 12/31/15. 
\textsuperscript{126} https://www.nyse.com/publicdocs/nyse/data/NYSE_Market_Data_Pricing.pdf 
\textsuperscript{127} While the ultimate disposition of this litigation is still pending, NetCoalition prevailed over the SEC in objecting to the fees and the decision by the DC circuit court. NetCoalition, Petitioner V. Securities and Exchange Commission, Respondent. NYSE ARCA, Inc. and The NASDAQ Stock Market, LLC, Intervenors, Nos. 09-1042, 09-1045, Decided: August 06, 2010. The matter is currently before an SEC administrative law judge.
Market data plan governance

Each of the NMS plans associated with the individual pools of market data are governed by an operating committee that has one representative from each SRO participant. There are no broker-dealer or bank committee members, nor are there any data vendors or members of the public on the operating committees of these plans. This is notable because market data are an essential input for equity trading, yet the trading public are not exchange members and thus have no direct representation in plan governance. The current market data regime was established when exchanges were closer to public utilities than they are now, before they became public for-profit entities. The SRO members of the operating committee, as producers of market data, stand to benefit from approving higher fees. Somewhat surprising from a governance perspective are the voting requirements of the plans. Making changes to the existing plans, including any changes to the fee structure of the plans, requires a unanimous vote of all participants. The SEC has issued a concept release in 1999 touching upon these issues, but as yet little has changed in plan governance.128

Concerns go beyond simply the voting protocols of the plan operating committees. Some critics have observed that the exchange personnel who are responsible for marketing consolidated core data are also responsible for the marketing of non-core proprietary data.129 This provides little incentive to improve core data, for which pricing is tightly regulated, when market participants may be willing to substitute more expensive non-core data in its place. Others are concerned with the reliability of the SIPS, the quality of data they produce, and the

129 See IEX Letter.
lack of transparency around the investigation of certain SIP shortcomings.\textsuperscript{130} Regardless of the
details, it would seem that a quasi-utility responsible for the provision of an input as essential to
our secondary markets as market data should, as a part of its formal governance process,
incorporate input from the public and other core constituencies of their markets.

\textbf{Market data and the oversight function}

Market data are important not only for the trading decision but also for the ex-post regulatory oversight function. Today, use of core and even most non-core data are insufficient to properly surveil technologically advanced trading firms. Market data identify transactions that occurred and quotations for orders that were placed, but do not attribute these trades or quotations to any particular party. More importantly, they cannot be used to identify where a particular order has been routed on its way to an eventual fill or cancellation. Regulators have the ability to access internal records of firms, but firms do not have obligation to make and retain the type of detailed audit trails necessary to reconstruct the trading process. Without such ability to reconstruct trading, surveillance of broker-dealers for enforcement or examination purposes is lacking. The ability to enforce and examine is essential for deterring bad actors and for fostering compliance by the regulated entity with applicable rules and regulations.

As a remedy to the situation, the SEC has approved a plan for the creation of a Consolidated Audit Trail (CAT).\textsuperscript{131} The CAT would allow tracking of order information and executions throughout the lifecycle of the order, including origination, routing, modification or cancellation, and execution. The audit trail would identify not only the broker submitting the

\textsuperscript{130} See SIFMA UTP letter.
\textsuperscript{131} Joint Industry Plan; Order Approving the National Market System Plan Governing the Consolidated Audit Trail, SEC Release No. 34-79318, November 15, 2016.
order but also would contain a unique identifier for each customer as well. Timely submission of such information would be required of broker-dealers. This information will markedly improve regulators’ ability to surveil for improper trading practices.

While this is an important piece of the solution to the oversight problem, it is not by itself likely to be enough. Because most trading today is algorithmic, the logic for actions taken on orders is embedded in the code resident at the broker-dealer. The code is the intellectual property of the broker-dealer, and these firms are very reticent to part with their intellectual property when regulators make investigative requests. Because secondary trading is so competitive, if any proprietary knowledge is leaked outside of their firms, it could easily become valueless since knowledge or use by other traders competes away any advantages conferred by the code. When the Commodity Futures Trading Commission (CFTC) proposed its new Regulation AT, which among other things allowed the regulator access the algorithmic trading code of its traders, the industry strenuously objected to overreaching by the CFTC and the potential compromising of valuable intellectual property. Yet without access to such code, it becomes problematic for a regulator such as the SEC to perform some of its basic oversight functions, including the deterrence of market manipulation, front-running, spoofing, or layering, most of which are based on the concept of fraud. Finally, it is worth noting that there are no requirements for broker-dealers to make and retain records of the operational versions of the order-routing and execution software. Because this software often evolves on a nearly continuous basis, as a practical matter it makes ex-post auditing of firm behavior a problematic exercise.

---

Validity of market data for best execution purposes

As discussed above, broker-dealers owe customers a duty of best execution when handling and routing customer orders. Most broker-dealers conceive of best execution as a process that seeks to obtain the best terms of trade for their customers. If conceived solely as a process, best execution obligations are relatively easy to describe and to surveil. As a hypothetical example, the best execution process for a particular broker-dealer for executing marketable customer orders may be to route the order to the market that shows the best price for the stock. In cases where there is a tie for best price, the broker-dealer would route to the market with the best price and greatest depth. This best execution process is easy to describe and relatively easy to audit on an ex-post basis. An examiner can, in principle, forensically rely on public quotation data to confirm that all outbound routes by the broker-dealer were to markets at the NBBO.  

More nuanced than the question of whether the broker-dealer followed their own prescribed best execution process is the question of whether the prescribed process was likely to obtain best execution for the client orders. This question is much more difficult to answer. We know the prints the broker received customer orders, but what we would most like to have to judge the efficacy of the broker’s best execution process are prints the broker would have received had the orders been routed in another manner, knowledge that is impossible to ascertain with precision. This lies at the heart of FINRA’s requirement that broker-dealers “regularly and rigorously examine execution quality likely to be obtained from the different markets trading a security.” In conducting such an analysis, and even in routing orders for actual execution,

\[133\] Even a simple exercise such as this may be challenging in a world of microsecond routing sequences, time-varying latencies, and imperfect clock synchronization.

broker-dealers must form judgments about the reliability of market quotations and the degree to which a customer order routed to a particular market center can be reasonably assured of executing at a favorable price. Publicly disseminated quotations are noisy signals of prices that would be obtained upon routing, not previews of the actual execution price itself. Prices may be better than, equal to, or worse than the quoted price at a given market center at the moment an order is routed. The differences can arise for a number of reasons. There may be hidden liquidity that would improve the execution of the marketable order. Or as discussed in O’Hara (2010) the quote may be flickering and be canceled before the order arrives at the market, which would harm execution quality. Or the state of the market may change with the arrival of a new better-priced order or the execution against the current quote of a quicker order. Sophisticated traders build complex proprietary models that assign probabilities to the likelihood of such events.

In establishing a benchmark for analyzing best execution, regulators must make judgments about the reliability of information in the quote record when assessing a broker-dealer’s performance. The record is often insufficient to reliably conduct such assessments. For example, suppose a broker-dealer internalizes an order for execution based on the price of a market whose quote at the time of in execution is inferior to the NBBO. The broker does so because they consider the superior NBBO quote short-lived and not reachable. Can such a broker be deemed to have violated their duty of best execution by not internalizing the customer order at the NBBO prevailing at the time the order was executed? Similarly, suppose a broker-dealer has the opportunity to route a customer buy order to a venue that will execute the order at a price of $10.00 with certainty. They also have the opportunity to route to another venue that has a 60% probability of executing at a price of $9.99, in a 40% chance of executing at a price of
$10.01. If the broker-dealer internalizes the order at $10.00 per share, selling the customer shares out of its own inventory, and simultaneously routes its own offsetting order to the 60/40 market and happens to buy the shares at price of $9.99, can the broker-dealer be judged to have violated its duty of best execution on an order-by-order basis for not internalizing the original customer order at a price of $9.99? The answer in part lies in the regulators belief about the reliability of signals contained in market data and, we believe, remains an open question.

The obligation of broker-dealers to purchase non-core data

As discussed in the best execution section above, broker-dealers who execute or route customer orders have an obligation to purchase core market data consisting of trades and quotations from all public exchanges. The same obligations do not apply to the use of enhanced non-core market data delivered via low latency direct feeds. As shown by Ding, et al. (2014), latencies between core and non-core market data can run up to 2 milliseconds. During times of high traffic volume, these latencies can be even larger and demonstrate considerable variation. Given that exchanges have latencies measured in the hundreds of microseconds, the best representation of what price a trader would get at a particular public exchange is probably more accurately represented by the direct feeds rather than by SIP data.

This causes a problem for a broker-dealer seeking to fulfill their obligation of best execution. If the broker uses only SIP data, they can expect they will be routing sub-optimally and will lose out to better informed traders making use of direct feeds. If they feel compelled to buy direct feed data, exchanges that understand the inelastic demand on the part of broker-dealers may raise prices on proprietary data knowing that broker-dealers will be forced to buy the data regardless of its cost. Broker-dealers are sensitive to this issue since they feel the
exchanges selling market data are charging them for a compilation of their own past trading history. To date the SEC has sidestepped this issue by not requiring purchase of non-core data, but this is less likely to remain the case as data from SIPs lose their relevance in the domain of high-frequency traders and market makers.

Perhaps the problem can in part be addressed through disclosure. Executing broker-dealers can be required to disclose their policies for obtaining and using the various types of market data. In this way customers can make informed choices about whether they want to pay the additional costs that may be charged by broker-dealers who consume non-core data. In principle, this same logic could apply to the relationship between introducing brokers and internalizing broker-dealers who execute their orders. If the internalizing brokers must disclose in their contracts their policies toward obtaining and using the various types of market data, introducing brokers, who retain their obligations of best execution, can make informed judgments about whether they will fully discharge their best execution obligations when they route to brokers who consume only a subset of available market data. The ultimate solution remains unclear, but what is certain is that there will continue to be friction between exchanges who charge for market data and broker-dealers who must pay for the right to use data that they feel should belong to them.

IV Fixed income secondary trading markets

Though the focus of this paper is on equity markets, the domestic fixed income markets are both larger and in more need of market structure reform than their equity counterparts. Whereas the market capitalization of listed equity markets is about $26.5 billion, the corporate, asset-backed, mortgage, treasury, agency, and municipal bond market in aggregate totaled $37.1
billion.\textsuperscript{135} U.S. equities trade in an integrated system of 12 public exchanges, more than 30 ATSs, and a substantial number of broker-dealers, all integrated by low latency high-capacity data linkages and a uniform system of both pre-and post-trade transparency. In addition, U.S. exchange-traded equities have the benefit of being publicly registered firms for which all fundamental information is available freely in a timely manner on the SEC’s EDGAR system. Spreads in equities are measured in pennies and in many cases stocks trade in increments of less than a cent.

Contrast this with the corporate and municipal bond markets. Here trade occurs in an over-the-counter dealer market. Pre-trade transparency is very limited, provided only by a few private systems and a growing number of bond ATSs.\textsuperscript{136} There is no integrated public display of dealer bids and offers, only a collection of private subscription-based systems. When a subscriber to one of these systems wishes to sell a bond, the system disseminates a “bid wanted” list to other subscribers, who are given a 1- to 2-hour window to post bids for the CUSIPs advertised for sale. There are also internal systems used by broker-dealers who have retail clients. These private systems display a select inventory of bonds for sale to the broker-dealer’s own customers, and typically only advertise bonds that the dealer already has in inventory.

Over the last decade, post-trade transparency is much improved. FINRA has created the TRACE system for corporate bonds, and the Municipal Securities Rulemaking Board (MSRB) has created the EMMA system for municipal bonds.\textsuperscript{137} Both systems are public last-sale trade reporting systems that disseminate time, price, and quantity of bonds traded in near real time.

\textsuperscript{135} SIFMA US Fact Book, 2016, pp 33-34.
\textsuperscript{136} There are now 23 TRACE-eligible ATSs.
\textsuperscript{137} TRACE has also been expanded to include certain mortgage and agency debt, and in July 2017 FINRA members will be required to report their Treasury trades to TRACE.
While these systems are applauded by small investors and their advocates, both dealers and large institutional investors generally opposed their creation. The likely reason for this opposition is that TRACE and EMMA dissipated the information advantages large and sophisticated traders in the previously opaque market.

Even with these improvements, trading costs in corporate and municipal bond markets far exceed those in equity markets.\textsuperscript{138} Harris and Piwowar (2006) show that municipal bond trading costs decrease with trade size and are substantially more expensive than comparable dollar amounts of equity securities. They also show that trading costs rise with time to maturity and instrument complexity. Green, et al. (2007) show that dealers earn their greatest profit in the municipal bond market on the smallest trades, and attribute this profit to the exercise of market power, which they show decreases in trade size. Using MSRB regulatory data, Sirri (2013) studies the manner by which dealers intermediate trades between customers and shows that trading costs fall with increasing trade size and rise with the complexity of intermediation required. The study also shows that it can take many days to fully intermediate a large trade, and that costs are lower if the dealer can lay off their principal risk position quickly.

Goldstein, et al. (2006) study the effect of increased transparency in the corporate bond markets and find it has either a neutral or positive effect on liquidity. They show that spreads decline on bonds that trade in regimes of greater post-trade transparency, which they attribute to investors’ ability to negotiate better terms of trade once they have access to such data. Edwards, et al. (2007) show that transactions costs in the corporate bond market decrease with trade size, and increase with the complexity of the bond, such as nonstandard payment, call, and maturity

\textsuperscript{138} There is a large and growing literature on secondary market trading in fixed income instruments. The authors apologize for the space constraints that prevent us from appropriately citing to all the relevant work in this space.
features. Bessembinder, et al. (2006) find that corporate bond execution costs fell one-half when an early version of the TRACE reporting system was introduced.

There are several fundamental factors that contribute to the high cost of trading in fixed income instruments described above. One is the large number of separate offerings in each of these markets. Municipal bond markets alone have over one million different bond issues outstanding. Many of these bonds go weeks or months without a single trade, and the odds of a natural buyer and seller having coincident interest to transact is low. Contrast this to the equity market with less than ten thousand publicly-traded equities. Also, because bonds trade only in dealer markets, and there are hundreds of different dealers asynchronously trading the bonds, it can be hard for two investors to find each other even if they do have simultaneous demand for trade unless they are customers of the same dealer. Finally, the information environment for fixed income is very different than it is for publicly-traded equities. Municipal issuers are not required to file public offering documents, such as prospectuses, with the federal government, and what limited obligations there are for disclosure arise indirectly through SEC rules imposed on brokers who trade or offer these bonds.

Although there has been progress in improving fixed income market structure, much more can and needs to be accomplished. As described above, post-trade transparency has improved for both corporate and municipal securities. In addition, the SEC and FINRA have recently improved trade disclosure for both municipal and corporate bond trades. These new

---

139 U.S. Treasury markets are notable exception to the high cost of domestic bond trading. Fleming, Mizrach, and Nguyen (2017) find that Treasury bonds trade in a basis point or less, and two electronic ETSs, BrokerTec in eSpeed, have substantial market shares of Treasury volume.
140 See 1934 Act, Section 15B(d).
141 For example, see SEC Rule 15c2-12.
142 For a more detailed description, see FINRA Regulatory Notice 17-24 (describing changes to FINRA Rule 2232) and MSRB Regulatory Notice 2016-28 (describing changes to MSRB Rule G-15 and Rule G-30).
rules will require disclosure of the dealer’s mark-up or mark-down on the customer’s confirmation statement if the dealer both trades as principal with a retail investor and engages in one or more offsetting transactions on the same day. The mark-up/mark-down confirmation disclosure must be expressed both as a total dollar amount and as a percentage amount. Regulators hope is that such disclosure will make investors more sensitive to trading costs and increase cost-based competition among dealers.

Bond market participants are quick to point out that bonds are not equities, and it is incorrect to presume that the same market structures are optimal for both. That said, it is highly likely that some meaningful improvements can still be made to fixed income market structure. One suggestion by the Financial Economists Roundtables, described in Harris, et al. (2015), is an enhancement to pre-trade transparency by requiring the development facilities that would allow for the display and execution of priced customer orders. Creation of the facility would be mandated either by the SEC or by FINRA. Broker-dealers who accept or execute customer orders could be required to place such orders into these display and execution systems. In this way, dealers and their customers at other firms would more easily be able to interact with what had previously been captive order flow of investors’ broker-dealer. As broker obligations with respect to best execution evolve in the presence of such systems, retail investors would have a chance to transact at much better prices than in a pure dealer market. Coupled with a requirement against inter-positioning, such a system might increase the likelihood that two customers come together for a natural trade without the involvement of an intermediary dealer. Improving markets by fostering investor transactions without the involvement of a dealer is one of the basic precepts behind the creation of the National Market System for equities.
Other improvements may arise by rethinking the obligations on brokers and advisers with respect to the bonds they recommend to retail investors. Best execution only addresses the terms of trade with respect to the bond under consideration for trade. Municipal securities and high-grade corporate bonds have very low probabilities of default, and thus with respect to credit risk show a great degree of fungibility. Most retail investors, when buying a municipal bond, classify their purchase desires in terms of general characteristics the bond, such as the credit rating, tenor, and traits such as callability and type of obligation (general obligation vs revenue). Few investors would walk into their broker’s office armed with the specificity to demand purchase of Scarsdale, New York 7-year non-callable water bond. Given the enhanced disclosures around bond costs, and with the future availability of agency platforms as suggested by the Financial Economists Roundtable, regulators may be able to encourage financial advisors to take greater consideration of the all-in-costs associated with bonds currently displayed in the various agency execution platforms when making initial purchase recommendations to their investors. Such a practice would lead to better after-cost yields for investors.

Ang and Green (2011) offer a completely different approach to the fixed income market structure issue by focusing on mechanisms to lower the borrowing costs for states and municipalities. They observe that both issuers and investors pay unnecessary fees and transactions cost because of the poor liquidity and transparency in municipal markets. Tax considerations cause the market for municipal bonds to be dominated by retail investors. The Tower Amendment prohibits the federal government from compelling the production and disclosure of core information from bond issuers, such as annual reports and offering documents, leaving investors largely in dark about the bonds they are buying.143 The authors propose the

---

143 See 1934 Act, Section 15B(d).
establishment of a not-for-profit platform, which they call CommonMuni, that would centralize the production and dissemination of information about issuers and offerings. It would offer also offer advice to issuers to help lower their financing costs. The authors observe that because bond complexity reduces market competition, it is in the interests of issuers to work together to standardize and harmonize the types of municipal bonds they offer. With larger and more homogenous offerings, market liquidity should improve and bond ownership should broaden. This is consistent with the findings of Harris and Piwowar (2006), who find that trading costs of more complex bonds are higher than those with simpler terms. Ang and Green estimate that the CommonMuni platform could be developed for about $25 million.

Doubtless there are other sound proposals for improvements to corporate, municipal, and other fixed income markets. Relative to domestic equity markets, improvements should come with relative ease. We would encourage regulators to refocus attention on fixed income markets and continue the trend of the last decade in improving the structure of these markets and the quality of information surrounding them. Especially for the municipal markets, the primary beneficiaries for such improvements are the retail investors who are at the heart of the SEC’s traditional mission.

V Conclusion

The primary purpose of trading markets is to provide a mechanism whereby investors can allocate their monies among productive firms. Such investment simultaneously allows investors to save for various life-cycle needs while facilitating capital formation on the real side of the economy. As such, any regulatory system for secondary trading must jointly serve the needs of both investors and issuers. Corporations receive their monies in primary market transactions
involving the issuance of securities. What subsequently occurs in secondary markets primarily serves the needs of investors, though through the informational role of markets, the aggregation of information reflected in prices is a valuable signal in the capital allocation process.

All regulations and market rules involve potential trade-offs, such as between pricing efficiency and considerations of fairness. Thus, any regulator is put in the position of making trade-offs between various stakeholders and constituents in the trading process. Striking the correct balance in making these trade-offs is essential, and certainly one of the key requirements to achieve an appropriate balance requires the collection and analysis of high-quality empirical data. On this dimension, there is good reason to be optimistic about future rulemakings. Over the last 10 years, the SEC has shown an increased proclivity to use financial economics as an organic part of the rulemaking process. In the not too distant past, use of economics was relegated to the so-called “back end” of rule releases. Here economics was often used on a post hoc basis to justify rules that were crafted without the benefit of economic insight. Whether because of several high-profile losses in appellate courts, or pressure from Congress, the SEC seems to have changed their rulemaking processes. The old Office of Economic Analysis, typically staffed by 20 to 30 professionals, has evolved into the Division of Economic Risk and Analysis, whose 110+ professionals form an important group within the Commission. Although it is too soon to judge, we should be hopeful that this change is permanent and not merely a reaction to pressures from the courts or the Hill.

With respect to equity markets, retail investors have never before had such low trading costs or such access to markets. Not only are spreads for small trades extremely low, but commissions charged by discount brokers are often under $10. And for this low cost, the service received by retail investors is far from rock-bottom. By most metrics, U.S. equity markets are
liquid, transparent, efficient, and competitive. Yet at the same time, there is a perception, whether correct or not, by some market participants that the market is unfair or rigged, as highlighted in Michael Lewis’ bestselling book *Flash Boys*. Even if the unfairness is far less severe than portrayed in this book, perceptions of unfairness can do lasting damage to markets. Whether through the unwarranted involvement of Congress to cure the perceived unfairness, or the withdrawal of skeptical investors from market participation, perceptions of fairness and integrity must be carefully managed.

As we discussed, for some issues the data seem relatively clear even if counterintuitive. For example, with respect to high-frequency trading, academic studies generally indicate this activity improves liquidity and pricing efficiency. But that does not mean it does so in every situation and at all times. Though the evidence generally indicates that high-frequency traders transact against temporary price movements, there is still a legitimate concern about whether a market substantially composed of high-speed computer-driven traders is as robust and resilient to shocks as a market intermediated by humans. One thing is certain—we will not be going back to the days of human traders walking the floors of physical exchanges and carrying paper tickets in their hands as they search for a contraside for their order. By 2008, the NYSE was forced to admit that the traditional market model, with a human specialist on the exchange floor charged with both positive and negative obligations, was no longer viable in the face of competition from off-exchange electronic market makers. When the NYSE migrated their market model away from their traditional form, and the floor specialist gave up their agency obligation to the limit order book, it was clear we had reached the point where humans were largely out of the market making business.
The Committee on Capital Markets Regulation (CCMR) has produced a report containing a list of suggested changes to regulations governing secondary trading equity markets.\(^{144}\) We concur with the authors of that study in their support of the use of pilot programs and independent studies as the basis for sound regulatory proposals. Recently the SEC has made greater use of pilot programs to investigate the effect of rule changes on measures of market quality. This is a positive change, and one that should be encouraged. Pilots have been used to study the effects of the uptick rule, equity tick sizes, securities lending, option penny trading, and options position limits. The SEC has even discussed a pilot for maker-taker pricing. Whether these pilots will provide useful data for future rulemaking is a difficult question.

Notable in the CCMR report is the incremental nature of the regulatory changes suggested by the Committee. That is not to say these changes are unimportant, but it is interesting to note that this group did not call for wholesale changes into our market structure, such as the repeal of large components of Regulation NMS, or the return to secondary markets intermediated by humans. Nor did the Committee ask for trading to be forced out of dark venues and away from dark order forms. Rather their recommendations revolved around promulgating and enhancing policies that foster competition. This, too, is a framework with which we agree. The choices the U.S. has made to encourage intermarket competition, at times even at the expense of intra-market competition, have served us well. We feel regulators should continue along this general course. At times, it does demand certain compromises that favor one business model over another, such as in the case of allowing payment for order flow, or permitting exchanges to pay maker-taker fees. But we believe the benefits of robust competition, and the

---

innovation that comes with it, will stand secondary trading markets in good stead in the decades to come.
References


Yao, C., Y. Mao (2015). Tick size constraints, high frequency trading, and liquidity. Working paper, University of Warwick and University of Illinois at Urbana-Champaign.


Table 1: U.S. Holdings of Equities ($ Billions, Market Value)

<table>
<thead>
<tr>
<th>Year</th>
<th>Value</th>
<th>%</th>
<th>Value</th>
<th>%</th>
<th>Total Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>$6,704.4</td>
<td>43.1</td>
<td>$8,840.3</td>
<td>56.9</td>
<td>$15,544.7</td>
</tr>
<tr>
<td>2002</td>
<td>$5,072.6</td>
<td>41.0</td>
<td>$7,314.6</td>
<td>59.0</td>
<td>$12,387.2</td>
</tr>
<tr>
<td>2003</td>
<td>$6,662.9</td>
<td>40.2</td>
<td>$9,924.5</td>
<td>59.8</td>
<td>$16,587.4</td>
</tr>
<tr>
<td>2004</td>
<td>$7,352.8</td>
<td>38.9</td>
<td>$11,527.5</td>
<td>61.1</td>
<td>$18,880.3</td>
</tr>
<tr>
<td>2005</td>
<td>$8,014.0</td>
<td>38.9</td>
<td>$12,586.7</td>
<td>61.1</td>
<td>$20,600.6</td>
</tr>
<tr>
<td>2006</td>
<td>$9,931.2</td>
<td>41.2</td>
<td>$14,162.0</td>
<td>58.8</td>
<td>$24,093.2</td>
</tr>
<tr>
<td>2007</td>
<td>$9,726.1</td>
<td>38.4</td>
<td>$15,602.9</td>
<td>61.6</td>
<td>$25,329.0</td>
</tr>
<tr>
<td>2008</td>
<td>$5,406.9</td>
<td>35.5</td>
<td>$9,830.5</td>
<td>64.5</td>
<td>$15,237.4</td>
</tr>
<tr>
<td>2009</td>
<td>$7,034.3</td>
<td>35.5</td>
<td>$12,767.5</td>
<td>64.5</td>
<td>$19,801.8</td>
</tr>
<tr>
<td>2010</td>
<td>$8,450.7</td>
<td>36.4</td>
<td>$14,791.7</td>
<td>63.6</td>
<td>$23,242.4</td>
</tr>
<tr>
<td>2011</td>
<td>$8,069.9</td>
<td>36.0</td>
<td>$14,376.3</td>
<td>64.0</td>
<td>$22,446.2</td>
</tr>
<tr>
<td>2012</td>
<td>$9,401.4</td>
<td>36.4</td>
<td>$16,452.4</td>
<td>63.6</td>
<td>$25,853.8</td>
</tr>
<tr>
<td>2013</td>
<td>$12,545.3</td>
<td>37.4</td>
<td>$21,001.3</td>
<td>62.6</td>
<td>$33,546.6</td>
</tr>
<tr>
<td>2014</td>
<td>$13,883.1</td>
<td>37.8</td>
<td>$22,887.9</td>
<td>62.2</td>
<td>$36,771.0</td>
</tr>
<tr>
<td>2015</td>
<td>$13,310.9</td>
<td>37.3</td>
<td>$22,376.3</td>
<td>62.7</td>
<td>$35,687.2</td>
</tr>
</tbody>
</table>

Source: SIFMA Fact Book, 2016
Table 2: Average Daily Equity Trading Volumes (Matched Volume for 5 days ended March 14, 2017)

**Panel A: Exchange Volume (shares)**

<table>
<thead>
<tr>
<th>Exchange</th>
<th>Tape A</th>
<th>Tape B</th>
<th>Tape C</th>
<th>Total Market</th>
<th>% Market</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEX</td>
<td>82,120,148</td>
<td>20,238,082</td>
<td>39,812,731</td>
<td>142,170,961</td>
<td>2.10%</td>
</tr>
<tr>
<td>CHX</td>
<td>12,428,436</td>
<td>12,428,436</td>
<td>4,698,218</td>
<td>29,012,382</td>
<td>0.43%</td>
</tr>
<tr>
<td>NYSE (N)</td>
<td>828,539,051</td>
<td></td>
<td></td>
<td>828,539,051</td>
<td>12.23%</td>
</tr>
<tr>
<td>NYSE Area (P)</td>
<td>227,228,365</td>
<td>290,102,387</td>
<td>130,203,073</td>
<td>647,533,825</td>
<td>9.56%</td>
</tr>
<tr>
<td>NYSE MKT (A)</td>
<td></td>
<td>13,064,685</td>
<td></td>
<td>13,064,685</td>
<td>0.19%</td>
</tr>
<tr>
<td>EDGX (K)</td>
<td>187,149,315</td>
<td>97,019,632</td>
<td>153,584,295</td>
<td>437,753,242</td>
<td>6.46%</td>
</tr>
<tr>
<td>BATS BZX (Z)</td>
<td>211,968,076</td>
<td>89,546,514</td>
<td>106,248,730</td>
<td>407,763,320</td>
<td>6.02%</td>
</tr>
<tr>
<td>BATS BYX (Y)</td>
<td>161,487,158</td>
<td>67,687,777</td>
<td>81,979,228</td>
<td>311,154,164</td>
<td>4.59%</td>
</tr>
<tr>
<td>EDGA (J)</td>
<td>81,714,899</td>
<td>39,020,576</td>
<td>38,513,757</td>
<td>159,249,232</td>
<td>2.35%</td>
</tr>
<tr>
<td>Nasdaq (Q)</td>
<td>363,754,710</td>
<td>120,616,142</td>
<td>449,915,452</td>
<td>934,286,303</td>
<td>13.79%</td>
</tr>
<tr>
<td>Nasdaq BX</td>
<td>100,673,993</td>
<td>31,452,682</td>
<td>55,690,142</td>
<td>187,816,818</td>
<td>2.77%</td>
</tr>
<tr>
<td>Nasdaq PSX</td>
<td>21,771,133</td>
<td>21,101,586</td>
<td>19,606,067</td>
<td>62,478,786</td>
<td>0.92%</td>
</tr>
</tbody>
</table>

**Panel B: Trade Reporting Facility (TRF) Volume (shares)**

<table>
<thead>
<tr>
<th>Facility</th>
<th>Tape A</th>
<th>Tape B</th>
<th>Tape C</th>
<th>Total Market</th>
<th>% Market</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasdaq TRF</td>
<td>1,148,611,147</td>
<td>526,907,736</td>
<td>718,242,565</td>
<td>2,393,761,448</td>
<td>35.33%</td>
</tr>
<tr>
<td>NYSE TRF</td>
<td>104,533,901</td>
<td>46,432,750</td>
<td>68,965,564</td>
<td>219,932,215</td>
<td>3.25%</td>
</tr>
</tbody>
</table>

Table 3: Weekly trading volume of 31 ATS reporting to FINRA (for week ended February 20, 2017)

<table>
<thead>
<tr>
<th>ATS Name</th>
<th>Total share volume</th>
<th>Total number of trades</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUA AQUA</td>
<td>260,800</td>
<td>43</td>
</tr>
<tr>
<td>BCDX BARCLAYS DIRECTEX</td>
<td>24,000</td>
<td>1</td>
</tr>
<tr>
<td>BIDS BIDS TRADING</td>
<td>93,169,106</td>
<td>117,574</td>
</tr>
<tr>
<td>BLKX BLOCKCROSS</td>
<td>11,436,732</td>
<td>1,065</td>
</tr>
<tr>
<td>CBLC CITIBLOC</td>
<td>3,740,544</td>
<td>191</td>
</tr>
<tr>
<td>CROS CROSSFINDER</td>
<td>285,513,409</td>
<td>1,602,153</td>
</tr>
<tr>
<td>CXCX CITI CROSS</td>
<td>22,061,154</td>
<td>90,803</td>
</tr>
<tr>
<td>DBAX SUPERX</td>
<td>201,627,980</td>
<td>1,097,385</td>
</tr>
<tr>
<td>DLTA DEALERWEB</td>
<td>26,421,232</td>
<td>81</td>
</tr>
<tr>
<td>EBXL LEVEL ATS</td>
<td>115,950,700</td>
<td>595,467</td>
</tr>
<tr>
<td>IATS IBKR ATS</td>
<td>22,874,024</td>
<td>76,419</td>
</tr>
<tr>
<td>ICBX INSTINET CONTINUOUS BLOCK CROSSING SYSTEM</td>
<td>81,773,936</td>
<td>368,331</td>
</tr>
<tr>
<td>ITGP POS</td>
<td>75,242,400</td>
<td>273,823</td>
</tr>
<tr>
<td>JPMX JPM-X</td>
<td>156,304,206</td>
<td>718,785</td>
</tr>
<tr>
<td>KCGM KCG MATCHIT</td>
<td>85,096,689</td>
<td>482,795</td>
</tr>
<tr>
<td>LATS BARCLAYS ATS (&quot;LX&quot;)</td>
<td>107,455,347</td>
<td>552,078</td>
</tr>
<tr>
<td>LMNX LUMINEX TRADING &amp; ANALYTICS LLC</td>
<td>5,269,324</td>
<td>152</td>
</tr>
<tr>
<td>LQNA LIQUIDNET H2O</td>
<td>12,315,900</td>
<td>1,196</td>
</tr>
<tr>
<td>LQNT LIQUIDNET ATS</td>
<td>22,557,100</td>
<td>567</td>
</tr>
<tr>
<td>MLIX INSTINCT X</td>
<td>95,361,619</td>
<td>461,797</td>
</tr>
<tr>
<td>MSPL MS POOL (ATS-4)</td>
<td>152,590,410</td>
<td>622,063</td>
</tr>
<tr>
<td>MSRP MS RETAIL POOL (ATS-6)</td>
<td>4,810,500</td>
<td>24,949</td>
</tr>
<tr>
<td>MSTX MS TRAJECTORY CROSS (ATS-1)</td>
<td>50,942,000</td>
<td>283,909</td>
</tr>
<tr>
<td>NYFX MILLENNIUM</td>
<td>29,374,620</td>
<td>90,706</td>
</tr>
<tr>
<td>PDQX CODA MARKETS, INC.</td>
<td>14,955,671</td>
<td>75,334</td>
</tr>
<tr>
<td>SGMA SIGMA X</td>
<td>85,087,955</td>
<td>439,786</td>
</tr>
<tr>
<td>UBSA UBS ATS</td>
<td>458,740,223</td>
<td>2,721,500</td>
</tr>
<tr>
<td>USTK USTOCKTRADE SECURITIES, INC.</td>
<td>2,905</td>
<td>66</td>
</tr>
<tr>
<td>WDNX XE</td>
<td>2,119,846</td>
<td>1,701</td>
</tr>
<tr>
<td>XIST INSTINET CROSSING</td>
<td>13,725,139</td>
<td>2,849</td>
</tr>
<tr>
<td>XSTM CROSSSTREAM</td>
<td>52,606,856</td>
<td>149,673</td>
</tr>
</tbody>
</table>

Table 4: Routing venues and routing decisions for an introducing broker for NYSE-listed stocks

<table>
<thead>
<tr>
<th>Route Venue</th>
<th>Total %</th>
<th>Market %</th>
<th>Limit %</th>
<th>Other %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citadel Execution Services</td>
<td>28.19</td>
<td>35.87</td>
<td>9.12</td>
<td>32.71</td>
</tr>
<tr>
<td>KCG Americas LLC</td>
<td>20.79</td>
<td>30.46</td>
<td>5.47</td>
<td>22.79</td>
</tr>
<tr>
<td>NASDAQ</td>
<td>15.89</td>
<td>0.00</td>
<td>47.66</td>
<td>9.81</td>
</tr>
<tr>
<td>G1 Execution Services</td>
<td>11.07</td>
<td>15.99</td>
<td>2.90</td>
<td>12.26</td>
</tr>
<tr>
<td>BATS (EDGX)</td>
<td>11.02</td>
<td>0.00</td>
<td>31.69</td>
<td>7.38</td>
</tr>
<tr>
<td>Two Sigma Securities</td>
<td>7.33</td>
<td>4.64</td>
<td>0.88</td>
<td>11.31</td>
</tr>
<tr>
<td>UBS Securities LLC</td>
<td>3.74</td>
<td>6.50</td>
<td>1.16</td>
<td>3.55</td>
</tr>
<tr>
<td>Total %</td>
<td>100.00</td>
<td>24.58</td>
<td>22.45</td>
<td>52.97</td>
</tr>
</tbody>
</table>