UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

DIPARTIMENTO DI DIRITTO PRIVATO E CRITICA DEL DIRITTO

MASTER'S DEGREE IN ECONOMICS AND FINANCE

MASTER'S THESIS

"HIGH FREQUENCY TRADING: IMPACT ON CAPITAL MARKETS AND REGULATION"

RELATORE:

CH.MO PROF. LUPOI ALBERTO

LAUREANDO: FERREMI ETTORE

MATRICOLA N. 1207029

ANNO ACCADEMICO 2020 – 2021

Il candidato dichiara che il presente lavoro è originale e non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere.
Il candidato dichiara altresì che tutti i materiali utilizzati durante la preparazione dell'elaborato sono stati indicati nel testo e nella sezione "Riferimenti bibliografici" e che le eventuali citazioni testuali sono individuabili attraverso l'esplicito richiamo alla pubblicazione originale.

The candidate declares that the present work is original and has not already been submitted, totally or in part, for the purposes of attaining an academic degree in other Italian or foreign universities. The candidate also declares that all the materials used during the preparation of the thesis have been explicitly indicated in the text and in the section "Bibliographical references" and that any textual citations can be identified through an explicit reference to the original publication.

Firma dello studente

TABLE OF CONTENTS

INTRODUCTION

Over the past five decades, technological innovations have revolutionized the securities trading industry. Human-intermediated capital markets, where stock markets' members manually traded one another physically interacting on the floor of the exchange, have been progressively replaced by exchanges' matching engines, where trades are electronically executed. The drastic increase of US trading volume at the end of the 1960s shed light on the necessity to bring the electronic and IT technology innovations inside financial markets: the automation of securities trading would have improved the functioning of capital markets, enabling them to trade an ever larger trading volume without loss in efficiency.

The electronification of capital markets made less expensive the provision of securities trading services that were usually offered by a small group of stock exchanges dominated by the NYSE, as liquidity supply or market data provide; thus, from the 1970s, and in particular over the last 30 years, new electronic trading venues started to operate, contributing to the fragmentation of the capital market in many trading centers in competition for order flow. The automation and fragmentation of financial market have determined the rise of a new breed of traders: algorithmic traders (ATs).

ATs are traders whose trading activity, from the submission of the order through its execution, is carried out automatically, without or with very limited human intervention, through a computer algorithm that pre-sets all the relevant order's parameters. Furthermore, the acceleration of information flows through the cables connecting the fragmented financial landscape's trading venues has instigated traders to an arms race for speed, culminated with the birth of fast ATs, called High Frequency Traders (HFTs).

As ATs, HFTs use algorithmic trading techniques but differently from the formers the core business of HFTs is to implement ultra-fast trading strategies through the use of low-latency infrastructures and high message intraday rates. In a financial market where informations flow at an ever faster speed across traders' computers' screen, the sentence of the unscrupulous financier Gorder Gekko "Informations is the most valuable commodity I know" from the famous 1987 Oliver Stone movie "Wall Street" seems to be particularly true, since also a difference of a few milliseconds in trading speed can determine whether a trader can profitably trade or not.

The global dominance of ATs in market share has stimulated a debate among finance academics and regulators on the effects of this new type of traders, in particular when engaging in high frequency trading (HFT) activity, on financial markets; in particular, the debate has focused the attention on the consequences of HFT on market quality measures, as liquidity, volatility and market efficiency; furthermore, after the events of the May 6, 2010 Flash Crash, scholars and regulators have highlighted the risk connected with some HFT aggressive trading strategies for the stability and integrity of financial markets.

This thesis has the object to describe the phenomenon of high frequency trading, to analyse the consequences of HFT on financial markets quality, reviewing the academic literature on the topic, in particular pointing out how the findings of these studies are not unambiguously, since most of the effects of HFT depend on the particular trading strategy adopted, and finally to analyse the response of financial regulation in US and EU to the new challenges of high frequency trading.

The thesis is organized in the following way. Chapter 1 describes the technological evolution and fragmentation of capital markets that have led to the development of algorithmic trading and high frequency trading; thus, it describes the evolution of algorithmic trading and the characteristics of AT and HFT, with the identification problem of HFT activity.

Chapter 2 describes in detail the trading strategies adopted by high frequency traders, dividing them into six main categories: market making strategies, arbitrage strategies, structural strategies, directional strategies, ghost liquidity strategies and trading on news.

Chapter 3 analyses the effects of HFT on different financial market metrics, as liquidity, volatility and market efficiency and the risks connected with HFT with respect to financial stability and integrity, with a focus on the event of May 6, 2010 Flash Crash, providing relevant academic studies about HFT.

Chapter 4 analyses the development of financial regulation in US and EU in response to the risks connected with algorithmic trading and HFT for financial stability and integrity, also discussing potential rules and capital markets' structural reforms to limit HFT activity, reducing its negative consequences for financial markets.

1 THE RISE OF HIGH FREQUENCY TRADING

1.1 Electronification of capital market

1.1.1 From consolidation to fragmentation

According to the Security and Exchange Commission's (SEC) definition, exchanges represent "any organization, association, or group of persons [...] which constitutes, maintains or provides a market place or facilities for bringing together purchasers and sellers of securities or for otherwise performing with respect to securities"[1]; operators managing stock exchanges' core business is to match quickly and cheaply buying and selling interests, and until 1970s consolidation of trading within a small number of stock exchanges was the tool used by US financial market regulators to help exchanges to pursue this purpose.

The rationale of market consolidation lies on the capacity of exchanges to attract more traders to their floor: the more traders an exchange can attract, the more easily they can trade with each other. For an exchange, large network of traders means more revenues from transaction fees; a solid profit margin allows them to charge lower fees on transactions, which leads to attracting even more traders to their floor**,** fueling this growth cycle further. The consolidation of ever higher trading volume improves the liquidity of trading venues, allowing traders to trade quickly and cost effectively; the resulting liquidity improvement also reduces dealers' inventory risk, attracting them to make markets on those exchanges and helping them to become even more efficient, offering a ready and reliable counterparty for investors. Furthermore, these extremely liquid markets represent a lucrative source of profit for dealers: they can gain the bid-ask spread from the immediacy they provide to impatient liquidity demanders on a large number of trades.

Since informed traders represent a small fraction of traders operating in financial markets, when an exchange attracts more traders, the fraction of informed traders as a proportion of the total number of traders should fall; this dynamic provides an incentive to informed traders to enter trading venues, since they will be able to increase their gains trading with a large number of uninformed traders. Moreover, since dealers are usually uninformed traders, a large number of uninformed traders allow them to partly reduce the adverse selection risk they are exposed to trading with informed traders.[2]

From aggregating the viewpoints of a multiplicity of traders, prices will reflect a larger quantity of available informations, improving market informational efficiency and functioning as a

---

[1] 15 U.S. Code § 78(c)(a)(1)

[2] Ananth Madhavan, 2000, Market Microstructure: A Survey, Journal of Financial Markets, Vol.3, issue 3, 205-258.

window into capital allocative efficiency. [3]

Despite the aforementioned benefits of a consolidated financial market, financial regulators found that a more concentrated trading activity within a few stock exchanges led to an impairment of capital markets' efficiency. Consolidation encouraged oligopoly of exchanges: they charged investors higher fees, provided weak infrastructures and delivered poor services. On the light of these concerns, of particular interest was the litigation emerged out of the infamous Nasdaq "odd-eights" scandal of the early 1990s. Christie and Schultz (1994) discovered that in a sample of 100 most liquid NASDAQ listed stocks, 60 of them rarely, if ever, were quoted in odd eights, but only in even eights; this suggested that the narrowest possible spread, that was equal to $0.125, was being prevented by the collusive behavior of NASDAQ dealers.[4] US regulators tried to solve these problems enhancing the competition between the existing exchanges promoting other trading venues to compete against them, such as Alternative Trading Systems; the fragmentation process of the US capital market began in the 1970s and throughout several regulatory changing ended up with the emanation of the Regulation National Market System in 2005.

1.1.2 The National Market System

In 1971, the SEC transmitted to US Congress its Institutional Investor Study on financial markets, suggesting a drastical change of capital market's structure; in this report, SEC proposed the creation of a central market system for securities of national importance, where all the bids and offers should have been reported in a consolidated manner, even if they are routed to different exchanges, with the main object to increase the competition in a market dominated by the incumbent NYSE and Nasdaq. At the end of 1972, with a market capitalization of $887 billions, NYSE counted for three quarters of the overall US market capitalization[5].

In the same year also another important report, the Martin Report, from the economist who released it, William McChesney Martin, proposed the same structural changes, pointing out the main objects the capital market reform should have pursued: the creation of a national securities system, instead of a market with local separated exchanges dominated by the NYSE, and the establishment of a uniform regulation among the various exchanges, limiting their self-

[3] Eugene F. Fama, 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, The Journal of Finance, Vol. 25, No 2, 383-417.

[4] William G. Christie, Paul H. Schultz, 1994, Why Do NASDAQ Market Makers Avoid Odd-Eighth Quotes?, Journal of Finance, Vol. 48, No 5, 1813-1840.

[5] Donald E. Farrar, 1974, Toward a central Market System: Wall Street's slow retreat into the future, in Journal of Financial and Quantitative Analysis, Vol. 9, No 5, 815-827.

regulation[6].

This reform would have allowed investors to search for the national best available price, increasing the competition between the exchanges. Technological innovations helped regulators to implement this reform; SEC proposed the creation of an electronic consolidated tape system recording all quotes and trade transactions in each listed security, that was able to connect all the existing trading venues.

The need of a financial market reform was also due to the technological progress that regulators recognized as of fundamental importance for the capital market efficiency, in particular after the so called “Back Room Crisis”; at the end of 1960s, trading volume in shares in the NYSE increased exponentially at a time when the mechanism for settling and clearing trades still required the physical transfer of related certificates at the end of the trading session; this physical process broke down, trades began to fail in extraordinary numbers and the loss of control over securities invited massive theft (between 1969 and 1970 New York Police Department and FBI estimated that almost $400 millions in NYSE financial transactions were stolen or lost)[7].

Congress responded to the SEC’s report with the Securities Acts Amendments of 1975, directing the SEC to establish the National Market System for the trading of securities; the main goal of the 1975 amendments was clearly enunciated in the added section 11A of the Security and Exchange Act, which contains an explicit statutory commitment to the establishment of a “national market system”; in particular, Congress stated that “it is in the public interest and appropriate for the protection of investors and the maintenance of fair and orderly markets to assure:

1. Economically efficient execution of securities transactions;
2. Fair competition among brokers and dealers, among exchanges markets, and between exchange markets and markets other than exchange markets;
3. The availability to brokers, dealers and investors of information with respect to quotations for and transactions in securities;
4. The praticability of brokers’ executing investors’ orders in the best market;
5. The opportunity, consistent with the other goals, for investors to execute orders without the participation of a dealer”[8].

[6] Martin W. McChesney, 1971, The Securities Markets: A report with Recommendations, in Virginia Law Review, vol 51, No 7.
[7] Blume Marshall E., Siegel Jeremy J., Rottenberd D., 1993, Revolution on Wall Street: The Rise and the Decline of the New York Stock Exchange, W.W. Norton & Company, New York.
[8] Securities and Exchange Act, § 11A(a)(1)(C), 15 U.S.C. § 78k-1(a)(1)(C).

At the light of the Congress directive, Regulation National Market System's most important laws were the Order Protection Rule, the Access Rule and the market data reform on the trading data dissemination and consolidation.

1.1.2.1 Regulatory implementation of the 1975 Amendments

The SEC's early efforts to establish a National Market System focuses on information linkages. The SEC regulates the dissemination and consolidation of quotations in NMS securities with the adoption of Rule 602, requiring national securities exchanges and national securities associations to "establish and maintain procedures and mechanisms for collecting bids, offers, quotation sizes, and aggregate quotation sizes from responsible broker or dealer who are members of such exchange or association".[9] In particular, national securities exchanges and national securities associations have to process the collected bids, offers and sizes in order to make available to all market participants "the best bid, the best offer, and aggregate quotation sizes for each subject security".[10] Furthermore, Rule 603 establishes the consolidation of informations in NMS securities from all the trading venues where they are traded through a Security Information Processor (SIP): "Every national securities exchange on which an NMS stock is traded and national securities association shall act jointly pursuant to one or more effective national market system plans to disseminate consolidated informations, including a national best bid and national best offer, on quotations for and transactions in NMS stocks"[11]Thus, the consolidation of NMS securities' NBB and NBO from all the trading venues that trade them enables the SIP to compute the National Best Bid and Offer (NBBO).

The most notable new rule is Rule 611, or in the SEC's terminology the Order Protection Rule, which requires that "A trading center shall establish, maintain and enforce written policies and procedures that are reasonably designed to prevent trade-throughs on that trading center of protected quotations in NMS stocks".[12] This rule ensures that investors obtain the best possible price for a given trade preventing trade throughs, that is the execution of a sell or buy order for an NMS security at a price respectively lower than a protected bid and higher than a protected offer, where protected bid and protected offer are quotations in an NMS stock that: " (i) is displayed by an automated trading center; (ii) is disseminated pursuant to an effective national market system plan; and (ii) is an automated quotation that is the best bid and the best offer of

[9] 17 C.F.R. §242.602(a)(1).

[10] 17 C.F.R. §242.602(a)(1)(i).
[11] 17 C.F.R. §242.603(a)(b).
[12] 17 C.F.R. §242.611(a)(1).

a national securities exchange or of a national securities association”[13]

SEC also regulates the access to quotations in NMS stocks; Rule 610 prohibits a national securities exchange or national securities association to “impose unfairly discriminatory terms that prevent or inhibit any person from obtaining efficient access through a member of the national securities exchange or the national securities association to the quotation in a NMS displayed through its SRO trading facility”[14]. The same rule caps fees for access to quotations; in particular, Rule 603 states that “ (1)If the price of a protected quotation or other quotation is $1.00 or more, the fee or fees cannot exceed or accumulate to more than $0.003 per share; or (2) If the price of a protected quotation or other quotation is less than $1.00, the fee or fees cannot exceed or accumulate to more than 0.3% of the quotation price per share”.[15]

1.1.3 Securities Information Processor

Advances in informational technology allowed the aforementioned legal framework to efficiently work; the “best execution” of the orders is only possible if traders have access to all the quotations in NMS stocks from all the trading venues where they are traded. SEC needed the implementation of a technological infrastructure, called Security Information Processor (SIP), that engaged in the business of “collecting, processing, or preparing for distribution or publication, or assisting, participating in, or coordinating the distribution or publication of, information with respect to transactions in or quotations for any security or distributing or publishing on a current and continuing basis, information with respect to such transactions or quotations”.[16] Moreover, as stated by William A. Schreyer, former chairman of Merrill Lynch, the only communication link among existing market centers does not solve the problem of best execution and limit order protection, but the display system must be complemented by the execution capability of transactions among different trading venues[17].

In 1976 the NYSE implemented the SIP for NYSE listed securities trades through the adoption of the Consolidated Tape System, which displayed last sale data on those securities from NYSE, AMEX and a variety of regional exchanges; the Consolidated Quote System providing informations on the quotations for the same securities from the same trading venues was implemented in 1978. As a result, all the securities traded and quoted on NYSE, AMEX and

---

[13] 17 C.F.R. §242.600(b)(61).

[14] 17 C.F.R. §242.610(a). SRO trading facility is a facility operated by or on behalf of a national securities exchange or a national securities association that executes orders in a security or presents orders to members for execution.

[15] 17 C.F.R. §242.610(c).

[16] 15 U.S. Code § 78c(a)(22)(A)

[17] Report issued by Merryl Lynch Pierce Fenner & Smith Inc. on June 14 1979 to US subcommittee, Progress toward the development of a National Market System.

other regional venues were displayed to a central data stream which connected all the exchanges that traded those securities. Similarly, for the Nasdaq listed securities the Unlisted Trading Privileges (UTP) Trade Data Feed collecting the last trade informations for Nasdaq listed securities was implemented, while the UTP Quote Data Feed displayed the quotations from all venues trading Nasdaq listed securities[18].

1.1.4 Smart Order Routing

The first exercise in the direction of an automated system with execution capability came in 1978 with the Intermarket Trading System (ITS). The ITS created an electronic link between NYSE and other regional markets, allowing brokers to route market orders to the exchange offering the best price at the time of the order. In the same year also the Cincinnati Stock Exchange adopted an automated system for transactions, called National Stock Trading System (NSTS). In 1981 the ITS merged with NSTS and in 1983 ITS and NASD's Computer Assisted Execution System were linked, making the first-ever automated link between the listed and OTC stock communities and creating the platform for the NMS[19]. The national market system grew progressively, updated many times until the SEC Regulation NMS of 2005, where today the shares of almost 8.500 companies are quoted[20].

With the evolution of IT technology, financial markets implemented a Smart Order Routing (SOR), a system that is able to connect and analyze in real-time price and quantity of all the financial instruments traded in the trading venues where they are traded and, according to a set of rules, guarantee the execution of the orders at the best available price in the National Market System. To clarify how the Smart Order Routing works suppose that an investor submits a market buy order to buy 1000 stocks of Microsoft, a stock that is quoted in three different trading venues, for instance NYSE, Nasdaq and BATS; suppose that the best ask for Microsoft in NYSE is $20.05 with 500 stocks available; the second best ask is $20.06 with 500 stocks available. In Nasdaq the best ask is $20.04 with 500 stocks available, while the second best ask is $20.05 with 500 stocks available. Eventually, in BATS the best ask is $20.02 with 500 stocks available, the second best ask is $20.03 with 300 stocks available and the third best ask is $20.05 with 300 stocks available. The SOR connects all the trading venues and allow the investors to

---

[18] Market Fragmentation and Its Impact: a Historical Analysis of Market Structure Evolution in the United States, Europe, Australia, and Canada, August 2013, Aite Group.

[19] James L. Hamilton, 1987, Off-Board Trading of NYSE-Listed Stocks: The Effects of Deregulation and the National Market System, in the Journal of Finance, Vol 42, No. 5, 1331-1345.

[20] Office of Analytics and Research, Division of Trading and Markets, SEC, Empirical Analysis of Liquidity Demographics and Market Quality for NMS, April 2018.

send its buy order to the one offering the NBO, that in our case is BATS, where an ask of $20.02 is available; however, the size of the NBO is not enough to entirely execute the investor's order because at that price level there are only 500 stocks of Microsoft available; consequently, the other 300 stocks are executed at the second best ask in BATS at $20.03. Finally, the remaining 200 stocks are routed to Nasdaq where they are executed at the best ask of $20.04 avilable in that market. Thus, through the SOR, the investor is able to execute its order across different trading venues obtaining an average buy price of $20.027, that is a better price than what the trader would have obtained if his order had been executed entirely on one trading venue (for example, the best average price he would have obtained in BATS would have been $20.029).[21]

1.2 The rise of ECNs and the market fragmentation process

The progressive electronification of stock markets transformed the traditional way they worked, shifting from primarily floor-based trading pits to electronic marketplaces, where orders are generated, routed and executed automatically.

Nasdaq began trading in 1971 as the world's first electronic stock market; the NYSE launched its automated trading platform in 1976, the Designed Order Turnaround (DOT) system, followed by the Super-DOT system in 1984, which allowed the transmission of orders to buy and sell to specialists electronically; however, this electronic improvement did not create an electronic exchange, rather it was called computer assisted trading ("CAT"), to signify that the specialist was merely being aided by technology and not replaced by it.[22] In 1969, Institutional Networks Corporation (later Instinet) began operations as a fully automated trading system, becoming the first Electronic Communication Network (ECN); unlike Nasdaq and NYSE, where orders were electronically routed to dealers and specialists that then handle them for their execution, Instinet allowed investors to have a "Direct Electronic Access" directly trading one another without the intermediation of an exchange's member. During 1990s other ECNs were launched, such as Island, Brut, Bloomber Tradebook, Archipelago and Redibook, Next Trade and Strike Technology.

The proliferation of several ECNs during 1990s increased the competition for order flows with NYSE and Nasdaq; whereas NYSE market share did not significantly decrease, giving up to ECNs 7% of its market share at the end of 2005, ECNs made tremendous progress in penetrating

---

[21] Alfonso Puorro, 2013, High Frequency Trading: una panoramica, Questioni di Economia e Finanza n° 198, Banca d'Italia, 7.

[22] Jerry W. Markham & Daniel J. Harty, 2008, For Whom the Bell Tolls: The Demise of Exchange Trading Floors and the Growth of ECNs, Journal of Corporation Law, Vol. 33, No 4, 865-939.

NASDAQ and accounted for approximately 57% of the market share within NASDAQ by the end of 2005 (Figure 1).[23]

The market fragmentation process seemed to be coming to the end in 2005; in order to contrast the increasing market share of ECNs, Nasdaq and NYSE conducted acquisition programs to reclaim their market share. In early 2006 the NYSE joined forces with Archipelago Exchange, becoming NYSE Arca, while Nasdaq acquired INET (the ECN resulted from the merger between Istinet and Island in 2002).[24] The two mega mergers temporarily stop the market fragmentation process and the market dominance of ECNs in Nasdaq market abruptly ended since Nasdaq acquired the largest of the ECNs: as we can see from Figure 2, at the end of Q2 Nasdaq consolidated its market share to hold approximately 52% of the market share in trading volume while ECNs market share was reduced to 4% from the 57% of 2005.[25]

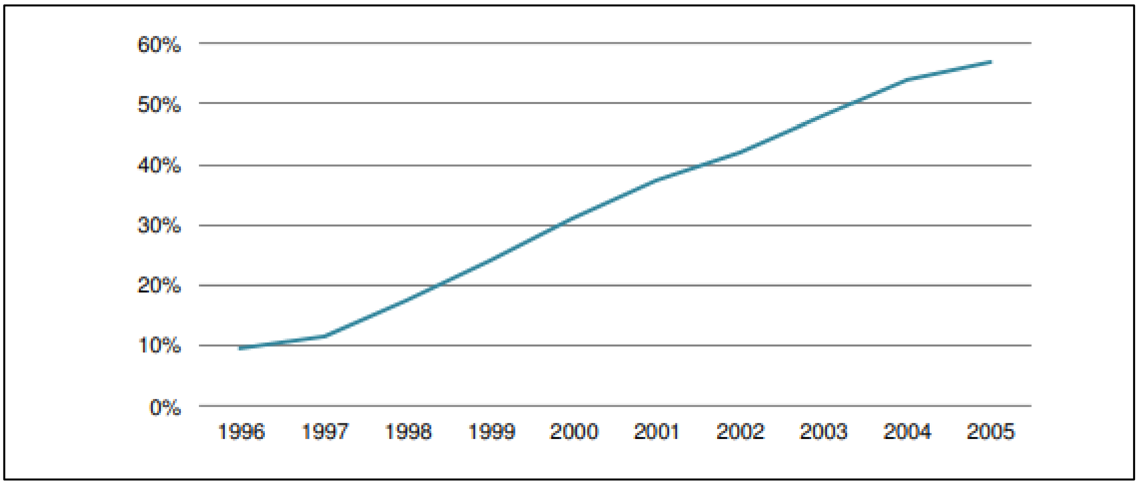


Figure 1: ECN penetration of Nasdaq, 1996 to 2005 (Market share based on average daily volume)

Source: Aite Group

[23] Id. 18

[24] Id.

[25] Id.

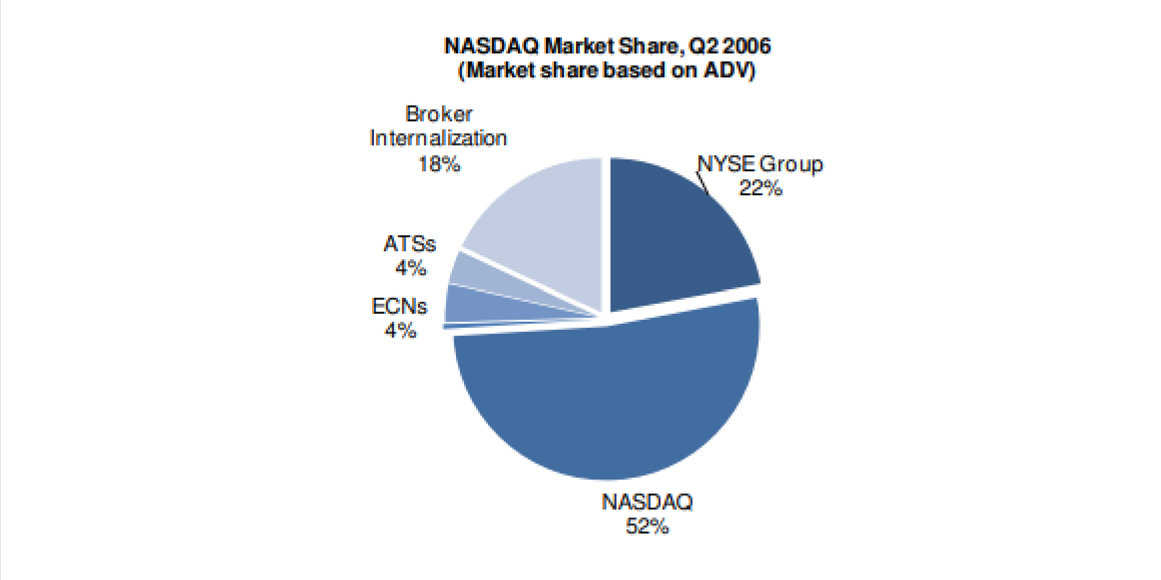


Figure 2: NASDAQ Market Share After Consolidation

Source: Aite Group

After the merger with Archipelago Exchange, NYSE Group owned 75% of market share, while ECN market share was reduced from 7% of 2000 to 1% of 2006. Eventually, in Q2 2006, Nasdaq and NYSE Group collectively accounted for 78% of the entire U.S. equities market, recreating the duopoly that had dominated the market before the rise of ECNs.

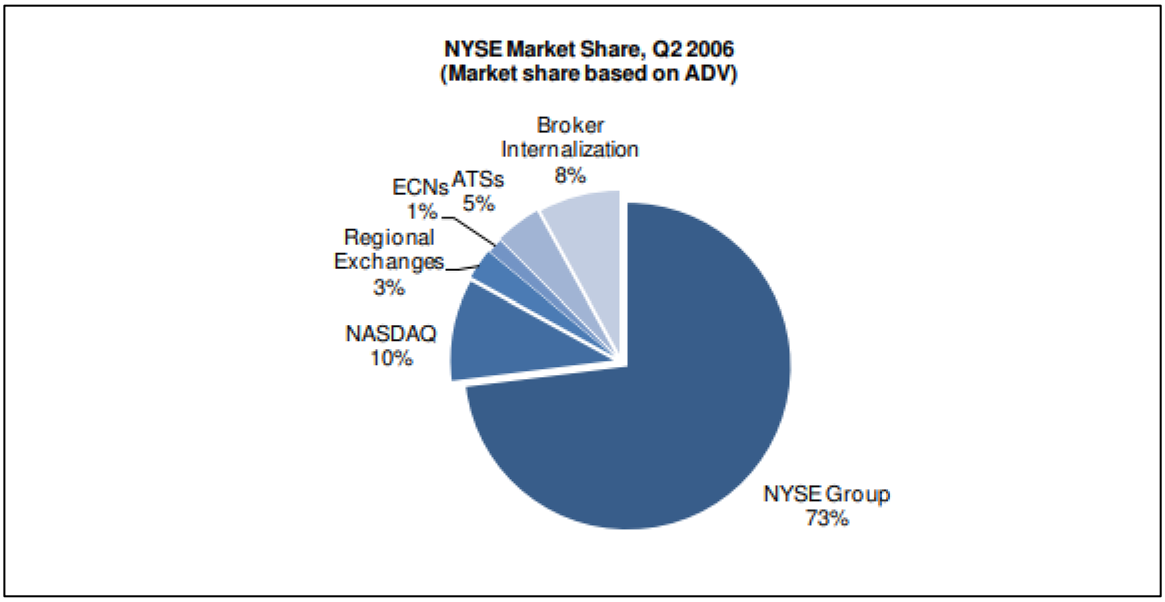


Figure 3: NYSE Market Share After Consolidation

Source: Aite Group

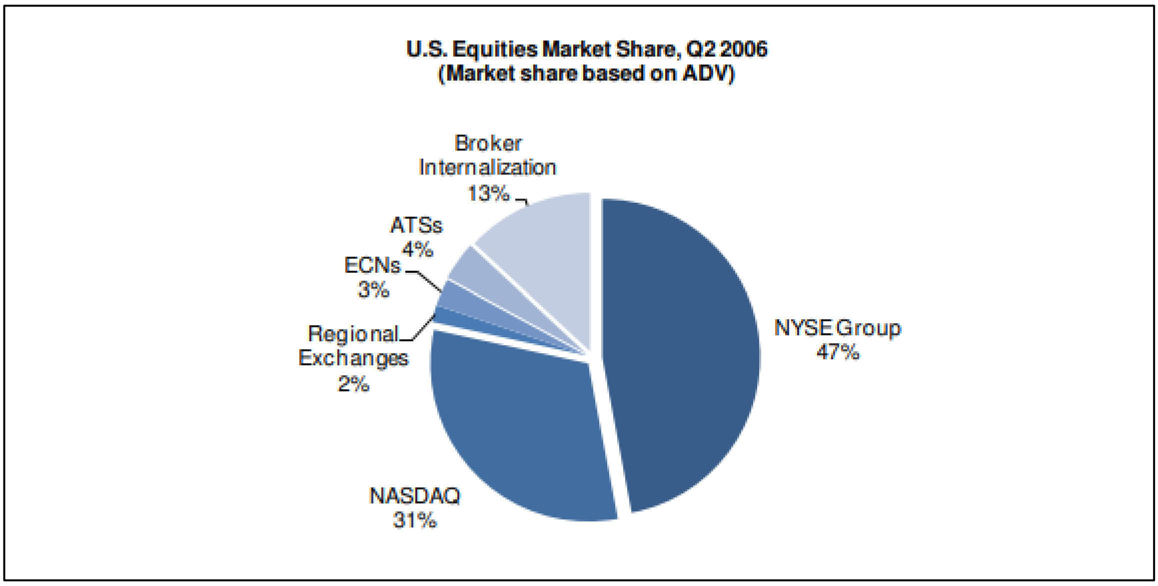


Figure 4: U.S. Equities Market Share Pre-Reg. NMS

Source: Aite Group

After the implementation of Reg NMS in 2005, US equities capital market knew another fragmentation process, mainly driven by broker-dealers which launched new light ECNs and dark pools to contrast the NYSE and Nasdaq duopoly. The two most important light ECNs was Direct Edge in 2009 and BATS in 2010, where main players are investment banks such as Goldman Sachs, Credit Suisse, Morgan Stanley. Even if dark pools were already launched before this period (e.g. Credit Suisse's Crossfinder in 2004)[26], the proliferation of dark pools began after Reg NMS.

The market fragmentation was also due to the fact that US exchanges began to operate more trading platform to test out new pricing scheme or launch new order types and products; for instance, Direct Edge manages EDGA, a low-cost exchange with a taker- maker pricing model, and EDGX, an exchange with a maker-taker pricing model offering high rebates for liquidity providers. In 2019 in US there were thirteen operating registered national securities exchanges, of which twelve are owned by three corporate entities, called "Exchange families".[27] Figure 5 shows how fragmented is the US market in NMS stocks among national securities exchanges (in terms of percentage of NMS stock trades, shares and dollar volume).

---

[26] Id.

[27] Report by the Staff of the U.S. Securities and Exchange Commission on Algorithmic Trading in U.S. Capital Markets, (August 5, 2020). The exchange families are (1) CBOE Global Markets, Inc., which owns CBOE BYX Exchange, Inc., CBOE BZX Exchange, Inc., CBOE EDGA Exchange, Inc., and CBOE EDGX Exchange, Inc.; (2) Nasdaq, Inc., which owns Nasdaq BX, Inc., Nasdaq PHLX LLC, and The Nasdaq Stock Market LLC; and (3) Intercontinental Exchange, Inc., which owns New York Stock Exchange LLC, NYSE Arca, Inc., NYSE American LLC, NYSE Chicago, Inc., and NYSE National, Inc.

Beside the proliferation of national securities exchanges, as we will see in more detail in the next section, from 1990s new trading center called Alternative Trading Systems, which includes the aforementioned ECNs, broker-dealer internalizers[28] and dark pools, significantly increase their participation in NMS stocks trading; in 2018 ATSs trading, also called off-exchange trading, represented the 35% of equity dollar volume in NMS stocks, while the 65% occurred on national securities exchanges (Table 1).

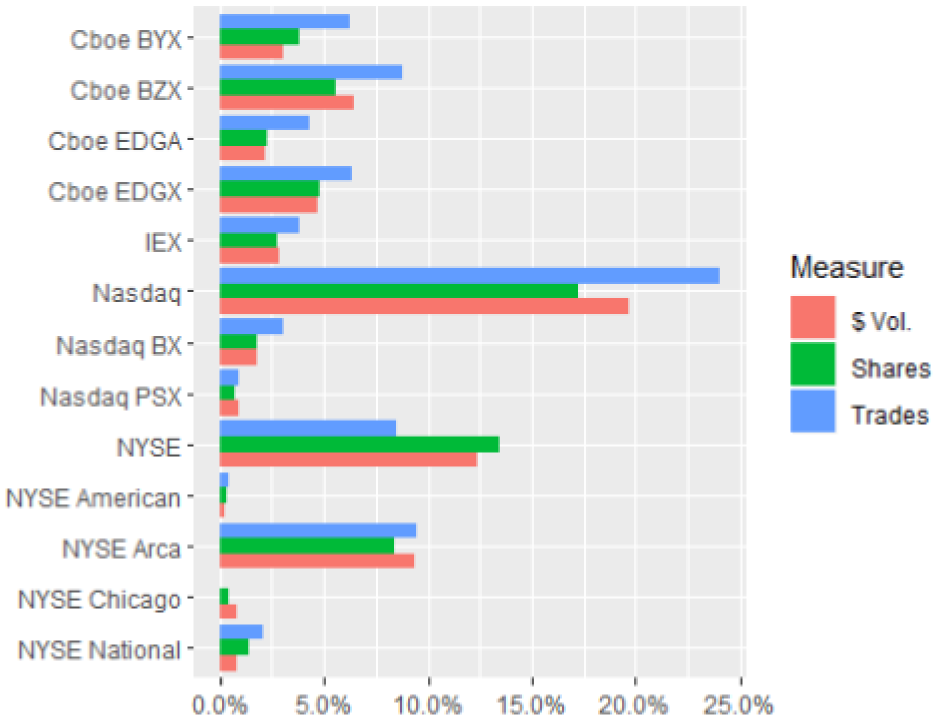


Figure 5: % of NMS Trades, Shares, and Dollar Volume in 2019

Source: NYSE TAQ

Venue	Trades	Shares	$ Vol.
Exchanges	78%	63%	65%
Off-Exchange	22%	37%	35%

Table 1: Percentage of all NMS stock trades, shares and dollar volume in 2018

Source: NYSE TAQ

[28] A broker-dealer internalizers executes the client's order out of its own inventory of security rather than routing it to an exchange or other platform in order to make money on the bid-ask spread; see U.S. Securities & Exchange Commission Investor Publications, Trade Execution: What Every Investor Should Know (Jan. 16, 2013), available at https://www.sec.gov/reportspubs/investorpublications/investorpubstradexechtm.html.

1.2.1 Alternative Trading Systems

ECNs were not recognized by the SEC as national securities exchanges or national securities associations: even if they met the SEC exchange definition facilitating to bring together buying and selling interests, they were not organized as exchanges or dealers' associations because they were not making a continuous market in securities. The 1975 amendments were drafted from the assumption that exchanges were self-regulatory organizations, with their members registered as broker-dealers[29]; conversely, ECNs did not involve broker-dealers in their platforms, where traders can access them without members' intermediation since an electronic order marching system automatically match limit orders.

The regulatory gap arisen with the birth of ECNs was eventually solved in 1998, when the SEC overhauled its rules relating to exchanges and other markets with the Regulation Alternative Trading System (ATS).[30] As a securities exchange or securities association, ATSs provide a place where investors can trade one another, but differently from the formers, they do not " Set rules governing the conduct of subscribers other than the conduct of such subscribers' trading on such organization, association, person, group of persons, or system"[31], where subscribers means any market participant that "has entered into a contractual agreement with an alternative trading system to access such alternative trading system for the purpose of effecting transactions in securities or submitting, disseminating, or displaying orders on such alternative trading system".[32] An ATS may include proprietary trading system, broker-dealer trading system and ECNs. According to Regulation ATS , ATSs have to register as broker-dealers[33]; in particular, any ATSs that have handled during at least 4 of the preceding 6 calendar months an average daily trading volume of 5 percent or more of the aggregate average daily share volume of an NMS stock, has to display the best bid and offer on that NMS stock to all the other NMS' exchanges and dealers' associations (the 5% rule).[34] Regulation ATS inaugurated a process of bringing ATSs into the national market system by bringing their best bids and offers into the public quote stream and giving the public the ability to execute against them. However, as we have seen in the previous section, currently in US equity market trades executed on ATSs, the so called off-exchange trading, do not display quotes: these venues are commonly referred as dark pools of liquidity. Although dark liquidity has long existed, for

[29] See, e.g., Exchange Act §6(b) (regulating the relationship between an exchange and its members); §6(c) (requiring that members be registered broker-dealers).
[30] Regulation of Exchanges and Alternative Trading Systems, Exchange Act Release No. 40760, 63 Fed. Reg. 70844 (1998).
[31] 17 CFR § 242.300(a)(1)(i).
[32] 17 CFR § 242.300(b).
[33] 17 CFR § 242.301(b)(1).
[34] 17 CFR § 242.301(b)(3)(B).

instance in the form of orders executed by upstairs trading desks or internalized by broker-dealers, dark venues significantly increased only with the growth of electronic trading.[35]

The growth of ATSs increased the competition with existing exchanges because these electronic platforms significantly decreased transaction costs and allowed market participants other than registered exchanges and associations to offer two services typically offered by the formers: market data provide and liquidity supply. Thus, with the rise of ATSs incumbent exchanges NYSE and Nasdaq lost their oligopoly on secondary markets for their listed securities. Investment banks such as Goldman Sachs or Merrill Lynch, that have always operated as brokers in regulated exchanges, began to internalize the orders of their clients, becoming broker-dealers internalizers, operating in the Over the Counter (OTC) market or using their own ECNs and competing with the exchanges for order flows. For example, Goldman Sachs owned Speer, Leeds&Kellogg, that is a specialist of the NYSE, than enabling the investment bank to supply liquidity, and a relevant stake in ArcaEx, an electronic stock and option exchange formed by a merger between NYSE and the Archipelago Exchange in 2006.[36]

The increased competition between incumbent exchanges and ATSs forced the formers to change their governance structure; historically, stock exchanges were organized as self-regulatory mutual association, where exchange's members paid an initial and annual member fee to rationalize the access to the market, where they made profits providing liquidity to investors acting as brokers and market makers. With the advent of automation in securities trading, the self-regulatory system was bound to fail; the homogeneity of interests between the exchange and its members that have make this system successful in the pre-automation era does not exist any longer. As we have seen with exchange members Goldman Sachs and Merrill Lynch, electronic trading significantly decreased the cost of providing liquidity, allowing them to compete directly with the exchanges of which they are members. Thus, the self-regulatory governance of stock exchanges shed light on a paradox: exchanges are regulated and regulate their competitors for listing and trading volumes.[37] Consequently, exchanges changed their governance structure through a process called demutualization; through this process, stock exchanges shifted from being non-for-profit self-regulatory cooperative system to for-profit publicly-traded firms, quoted on an exchange as any other listed firm[38]. The main function of demutualization was the separation between membership and ownership, which reduced the

[35] Issues raised by Dark Liquidity, Consultation Report, IOSCO (2010). Upstair trading refers to the situation where a listed stock is not executed in the listing exchange but buyers and sellers negotiate the price and conditions of the trade in the upstairs rooms of a brokerage firm.

[36] Macey, O'Hara, 2005, Markets to venues: Securities regulation in an evolving world, in Stanford Law Review, Vol 58, No. 2, 563-600.

[37] Id.

[38] Id:

control of intermediaries on the strategic positioning of the exchange: those that were members of the exchange ; furthermore, this process allowed exchanges to raise capital for expansion and technology innovations.[39]

Also listed firms promoted the proliferation of new trading venues, listing their stocks not only on the primary exchange where they first launched the IPO, But also on other trading venues (i.e. cross-listing), with the object to offer their investors different options for trade executions, such as narrow bid/ask spread or speed of execution. However, while for listing on different exchanges firms have to meet the listing requirements of each exchange, trading on ATSs does not require any listing requirements; in particular, ATSs can benefit from regulatory enforcement of exchanges where the stock they trade are listed without sharing the associated regulatory costs[40]. Furthermore, also investors and traders can trade on other trading venues without incurring regulation fees; thus, the regulation is a public good, and as other public goods, primary exchanges such as NYSE or Nasdaq can face free-rider problems from ATSs.[41]

1.3 Direct electronic access

1.3.1 Direct Market Access

The automation of trading system allows trading venues to handle an ever increasing volume of trades without loss of efficiency, as we have seen during the Back Room Crisis. Automated trading systems offer extremely high speed, or low-latency, order responses and executions, reducing the risk to trade at a stale price; ECNs further reduced latency by offering investors a Direct Electronic Access to the financial marketplace, enabling them to trade without the obligation to register as a member of the exchange. Direct access trading represents the shift in the access and control of trading from the sell-side to the buy-side of financial markets. There are two main types of DEA: the Direct Market Access (DMA), and the Sponsored Access (SA). With DMA, clients can use broker's infrastructures to send their orders to the trading center and control its execution, hence the moniker "Zero touch" order, since the broker is no longer involved in the trading process; using broker's exchange connectivity infrastructures, the broker has full control over the customer flow, including pre and post-trade compliance and reporting.

[39] Benn Steil, 2002, Changes in the Ownership and Governance of Securities Exchanges: Causes and Consequences, in Brookings-Wharton Papers on Financial Services, 61-91.
[40] Id. 35. In 2003 NYSE charged $113.506.000 in regulatory fees, 10% of exchange's revenue.
[41] Id.

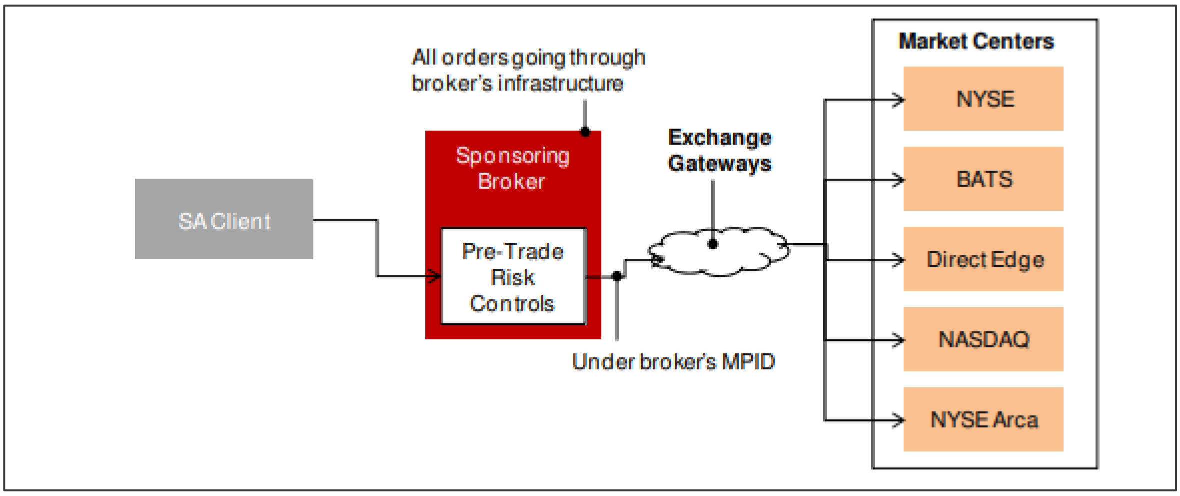


Figure 6: Direct Market Access Work Flow
Source: Aite Group

1.3.2 Sponsored Access

As DMA, Sponsored Access allows the client to connect to the market using their broker's market participant identification (MPID), but without having to go through their exchange infrastructure. The market participant that receives the sponsored access is called sponsored participant, while the broker-dealer who allows it to use its MPID is called sponsoring broker.[42] There are two specific types of Sponsored Access based on whether real time risk checks exist at an account level:

- Filtered sponsored Access: through this model, even if the sponsored participant does not use the broker's exchange infrastructure, sponsoring broker can still set up and monitor pre-trade risk parameters, and, if necessary, remotely modify or shut down trading activities before undesirable transactions have actually occurred[43].
- Unfiltered Sponsored Access (or naked access): under this sponsored model, as for the filtered one, the sponsored participant receives direct access to marketplace by an infrastructure provided by the sponsored broker but there is no pre-trade risk control by the sponsoring broker: it only receives a drop copy of the transaction by the client, which may not be received in real time and may not be used for risk management[44].

[42] Id 21.
[43] Id.
[44] Id.

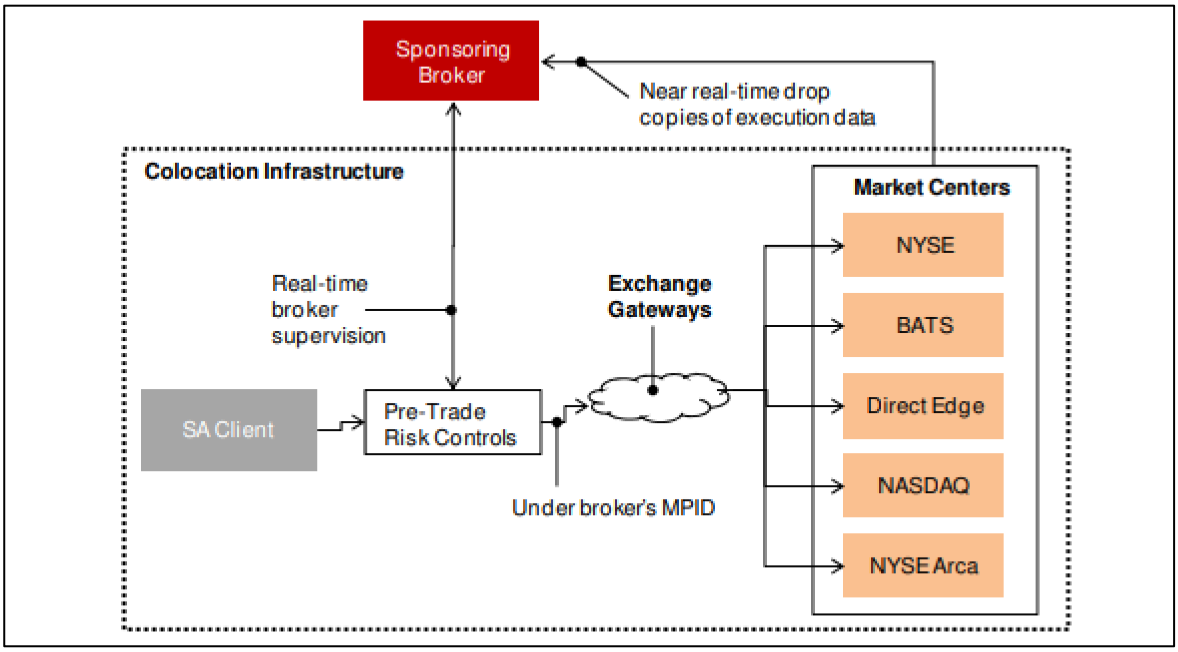


Figure 7: Filtered Sponsored Access Work Flow

Source: Aite Group

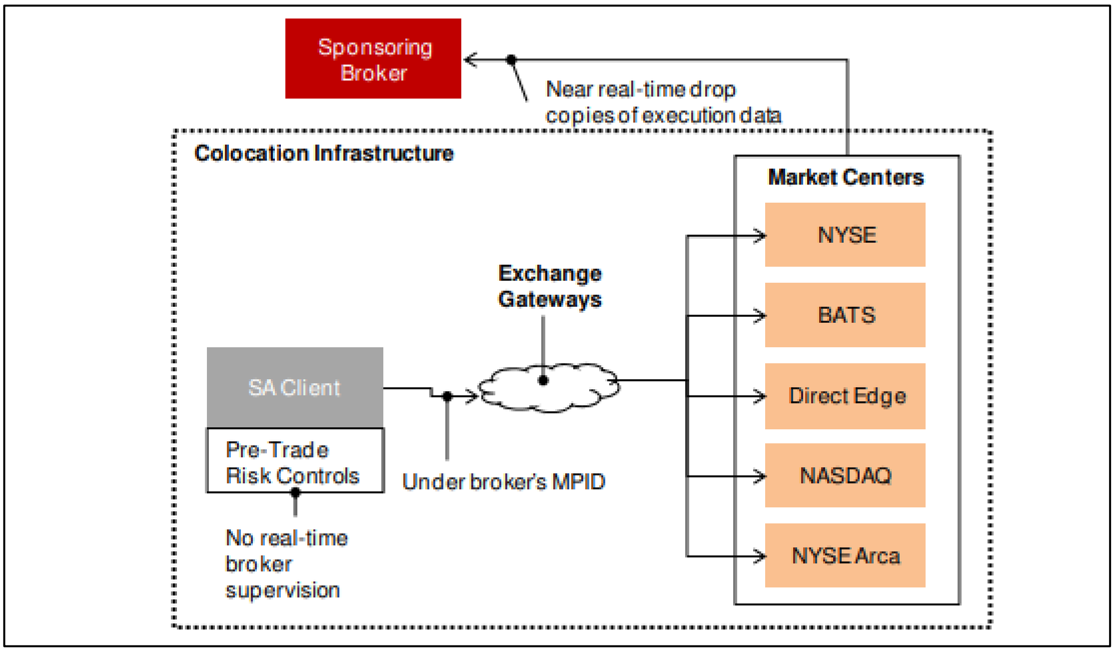


Figure 8: Unfiltered Sponsored Access Work Flow

Source: Aite Group

1.4 Financial market infrastructures

When a trade is executed through either the intermediation of a broker or via Direct Electronic Access, financial markets' infrastructures must guarantee the commitment to trade translates into an actual exchange between the buyer and the seller of a financial instrument; the effectiveness of this process is ensured by the supply of two post-trade services: clearing and settlement, respectively provided by Central Counterparties (CCPs) and Central Security Depositories (CSDs). The current clearing and settlement process was not built overnight, but it took years of studies and progressive adjustments that started with the Paperwork Crisis ("Back Room Crisis") of the end of 1960s, which shed light on the needs of a financial market structure reform that leveraged the automation in the securities industry to establish a uniform, coordinated and nationwide system for the clearing and settlement of financial transactions.
CCPs are the financial infrastructures that ensure a trade on an exchange is settled by interposing themselves between the buyer and the seller of a financial instrument: becoming the buyer of every seller and the seller of every buyer, CCPs mitigate the counterparty credit risk, that is the risk in a bilateral transaction that a party defaults on its obligations.[45]Through the clearing mechanism of CCPs, market participants do not have to bear the cost of assessing the creditworthiness of each trader with which they interact since the CCP will honor the obligation in the case of default of one party. The overall counterparty risk is also reduced by the netting, or off-setting, mechanism implemented by CCPs: instead of clearing each single trade a market participant executes with respect to a certain financial instrument, CCPs aggregate traders' positions on it to achieve a net position owed by the trader; for example, if a trader has a short position on a stock equal to $1200 and a long position on the same stock equal to $2000, instead of clearing each position, the CCP clears the trader net position, that in this case is equal to a net long position of $800.

CCPs act as a buffer against the systemic risk the default of a counterparty may cause, by absorbing the losses before they propagate across the financial system; however, even if CCPs reduce market participants' default risk, they do not eliminate it because CCPs themselves can fail if they are not adequately capitalized against market participants' default[46]. Thus, CCPs mutualize the losses from counterparty defaults among their clearing members, requiring them to post a margin as collateral for their exposures, that is an amount equal to a percentage of the clearing member's exposures to market participants, the so called initial margin.[47] The

[45] Guido Ferrarini, Paolo Saguato, 2014, Regulating Financial Market Infrastructures, ECGI Working Paper N°. 259.
[46] Yeshav Yadav, 2013, The Problematic Case of Clearinghouses in Complex Markets, Georgetown Law Journal, Vol. 101, 387-444.
[47] Id. 41.

mechanism of margins to mitigate counterparty risk is particularly important when CCPs clear transactions in derivative contracts, with the procedure known as "marking to market" or "daily settlement"; since in derivative markets the settlement of a transaction occurs on a future date, margin mechanism allows CCPs to daily settle gains and losses of a market participant on its derivative position, spreading the potential losses on it over the life of the derivative transaction instead of accumulating it until the expiration of the contract, which would increase the default probability of the counterparty[48]. To clarify how margins work we can consider a trader entering a long position on two December gold futures contracts on the COMEX division of the NYMEX; suppose that the current futures price is equal to $1.450 per ounce. If the contract size is 100 ounces, the trader has contracted to buy 200 ounces at this price in December, for a total amount of $ 290.000. the initial margin the CCP requires is equal to $6.000 per contract, or $12.000 in total. If at the end of the day the price of the futures is declined to $1.441 per ounce, the trader long in the two contracts has accrued a loss equal to $1800: the 200 ounces of December gold he has contracted to buy at $1450 at the end of the day can be sold for only $1441. Thus, at the end of the day the CCP will transfer this loss from the margin account of the long trader to the one of the short trader. If the trader continues to accumulate losses until the margin posted drops below the so called maintenance margin (the minimum margin level), it will trigger a margin call: the CCP will require the trader to reconstitute the initial margin, otherwise the CCP will close his position.[49]

Beside the margin mechanism, CCPs' financial soundness is also enforced by additional prudential requirements: clearing members must provide financial resources in the form of capital, default funds and collateral for their cleared transactions. The mutualization of losses after the default of a clearing member occurs through the so called "default waterfall" procedure, according to which losses are firstly covered by the margins of the defaulting members, then by their default fund contributions and finally by the default funds of non-defaulting members.[50]

In the wake of the 2008 financial crisis and after the G20 Pittsburgh summit on 25 September 2009, the clearing mechanism was extended to some classes of OTC derivatives with the aim to increase the post-trade transparency on a class of instruments that was blamed to have amplified the systemic risk because of the lack of regulatory oversight on these transactions; in

[48] While in derivative markets the settlement of a transaction can occur months after the trading day, for securities' transactions the settlement period is equal to 2 business days after the trading day; see 17 CFR § 240.15c6-1 for US regulation and Article 5(2) of Regulation (EU) No 909/2014 of the European Parliament and of the Council of 23 July 2014 for EU regulation.

[49] J.C. Hull, 2017, Options, Futures and Other Derivatives, 10th ed., Pearson.

[50] Id. 41.

particular, in US the 2010 Dodd-Frank Act and in Europe the adoption of the European Market Infrastructure Regulation (EMIR) in 2012 transferred respectively the trading of swaps and standardized OTC derivatives from OTC markets to regulated trading venues.[51]

Once the financial transaction is cleared it must be settled, which means that the financial instrument must be transferred to the buyer's account and the cash payment must be transferred to the seller's one; the settlement service is provided by Central Security Depositories (CSDs). Market participants use Custodian Banks to hold their assets and cash through safekeeping and cash accounts and Custodian Banks in turn hold their clients' assets in centralized custody via the CSDs, either in certificate form or dematerialized.[52] The interaction between sellers' and buyers' safekeeping and cash accounts is efficiently implemented by Security Settlement Systems (SSSs) through which Custodian Banks are interconnected. Each Custodian bank sends to the SSS settlement instructions, such as trade date, intended settlement date and counterparty bank's account details in order to deliver a quantity of securities from the account of the seller to the buyer's one and to instruct the payment of an amount of cash to the seller's account[53]. In particular, the transfer of the ownership of a security does not require the transfer of the associated physical certificate (one of the main causes of the market inefficiency that culminated with the Paperwork Crisis), but it can easily and quickly transferred through an electronic bookkeeping entry recorded on the CSD's account.

Nowadays, the US-based Depository Trust & Clearing corporation (DTCC) is the largest company that provides these services, with $2.150 trillions of US securities deals settled in 2020.[54] DTCC was born in 1999 with the merger between the Depository Trust Company (DTC), the largest CSD in the world, and the National Securities Clearing Corporation (NSCC), both constituted in the aftermath of the Paperwork Crisis[55].

The last post-trade service is the reporting and dissemination to public of transactions' data, which comprises price, volume and time at which the last trade has been executed. As we have introduced in section 1.1.3, nowadays in US there are two Security Information Processors: the Consolidated Tape Association's (CTA) feeds and the Unlisted Trading Privileges' (UTP) feeds. The CTA feeds are the Consolidated Quote System (CQS) and the Consolidated Tape System (CTS); the former reports top-of-book quotes and calculates the NBBO of NYSE LLC

[51] Id.

[52] Association for Financial Markets in Europe, The role of post-trade services in the financial sector, 2015.

[53] Id.

[54] Philip Stafford, US stock clearing house proposes quicker settlements after GameStop saga, February 24, 2021, https://www.ft.com/content/4c33198e-245d-4aaf-a106-86d0de4d7cca.

[55] DTC was founded in 1973 with the object to create a centralized deposit for all NYSE security certificates, in order to avoid that each single broker has to physically transfer the certificate from the seller to the buyer; in 1976 the NSCC was founded as a centralized clearing house to provide multilateral netting between NYSE market participants.

(Network A) and Bats, NYSE Arca, NYSE American and other regional exchanges (Network B) listed securities, while the latter reports the last sale price and volume of the same securities[56]. The UTP feeds are the UTP Quotation Data Feed (UQDF) and the UTP Trade Data Feed (UTDF), that are the analogous quote and trade feeds for Nasdaq listed-securities (Network C)[57].

Under EU regulation, trade data reporting services are provided by four entities: the trading venue, the Approved Reporting Mechanism (ARM), the Approved Publication Arrangement (APA) and the Consolidated Tape Provider (CTP). ARMs are persons authorized to provide, on behalf of investment firms, to National Competent Authorities or European Securities and Markets Authority (ESMA) details on investment firms' transactions, in order to help them to meet their obligations under article 26 of MIFIR, which requires them to provide transactions' details to competent authorities no later than the close of the following working day[58].

The APAs are persons authorized to make public trade data on behalf of investment firms, allowing them to meet their obligations to make public the volume, price and time at which trades have been concluded.[59]

Finally, CTPs are the analogous of US SIPs: they are persons authorized to collect trade data from regulated markets, MTFs, OTFs and APAs and to consolidate them into a continuous live data stream which provides real-time price and volume of trades in financial instruments.[60] As we have seen for the clearing process, also for post-trade data reporting US and Europe extended the regulation to OTC market transactions through the Trade Repositories (TRs), data warehouses which collect informations on all the derivative transactions occurred on trading venues or OTC.[61] By providing informations on OTC derivative transactions, TRs allow market regulators to monitor the actual size and distribution of exposures in derivative contracts, thereby supporting risk management, market discipline and effective oversight, regulation and

---

[56]See https://www.nyse.com/data/cta; The current Participants include the Cboe BYX Exchange, Inc., Cboe BZX Exchange, Inc., Cboe EDGA Exchange, Inc., Cboe EDGX Exchange, Inc., Cboe Exchange, Inc., Financial Industry Regulatory Authority, Inc., Investors Exchange LLC, Long-Term Stock Exchange, Inc., MEMX LLC, MIAX Pearl, LLC, Nasdaq BX, Inc., Nasdaq ISE, LLC, Nasdaq PHLX LLC, Nasdaq Stock Market LLC, New York Stock Exchange LLC, NYSE American LLC, NYSE Arca, Inc., NYSE Chicago, Inc., and NYSE National, Inc.

[57] See https://www.utpplan.com/participants for the list on UTP participants.

[58] Article 4(1)(54) Directive 2014/65/EU; MIFIR, Art. 26(1) Regulation 600/2014/EU

[59] Article 4(1)(52) Directive 2014/65/EU; Articles 20 and 21 Directive 2014/65/EU: Investment firms and Systemic internalizers have to report data on trades on equity financial instruments (shares, depositary receipts, ETFs, certificates and other similar financial instruments) and non-equity financial instruments (bonds, structured finance products, emission allowances and derivatives)

[60] Article 4(1)(53) Directive 2014/65/EU.

[61] Id. 36. 2010 Dodd-Frank Act extended TRs' registration to any swap contract, both cleared and uncleared (see §727 of the 2010 Dodd-Frank Act), while EMIR requires all counterparties and CCPs to report details of any derivative transactions, including any modification or termination, to a TR, see Article 9(1) Regulation 648/2012/EU.

supervision that can enforce market transparency and financial stability.

1.5 Algorithmic Trading

The rise of modern electronic markets has led to the birth of a new breed of traders: the algorithmic traders. MIFID II defines algorithmic trading as a computer-powered trading technique which, through pre-coded rules implements algorithms that "automatically determines individual parameters of orders such as whether to initiate the order, the timing, price or quantity of the order or how to manage the order after its submission, with limited or no human intervention" [62]. From this time on, all the trading activity can be divided into algorithmic trading and non-algorithmic trading, depending on whether or not traders use computer algorithms to make trading decision.

Specific characteristics of AT excluding HFT
1) Minimize market impact of block orders
2) Benchmark based strategy
3) Long term holding periods
4) Working an order through time and across markets

Table 2: Specific characteristics of AT excluding HFT

Source: P.Gomber, B.Arndt, M.Lutat, T.Uhle ,2011

1.5.1 High Frequency trading

The most important evolution of algorithmic trading is high frequency trading ("HFT"), that is a subset of algorithmic trading whose main features are:

- "the use of extraordinarily high speed and sophisticated computer programs for generating, routing and executing orders;
- the use of co-location services and individual data feeds offered by exchanges and third parties to minimize network and other types of latencies;

[62] MIFID II, Art. 4(39); the provision also specifies that algorithmic trading does not include systems that are only used for the purpose of routing orders to one or more trading venues or for the processing of orders involving no determination of any trading parameters or for the confirmation of orders or the post-trade processing of executed transactions.

- very short time frames for establishing and liquidating positions;
- the submission of numerous orders that are cancelled shortly after the submission;
- the closing day position as flat as possible"[63].

From the above characteristics of HFTs we can identify the main differences between algorithmic trading and HFT; the use of high-speed trading through the adoption of exchanges' and third parties' co-location services and direct market feeds allows HFTs to trade very high number of orders during a trading day with respect to non high frequency algorithmic traders. Usually, for each of the numerous trades executed, HFTs gain a very low margin and for this reason they execute large volume of trades to reach an overall significant margin. HFTs also leverage their trading speed to act as market makers, rapidly updating and cancelling limit orders: as we will see discussing about HFT identification, one of the most used measure to detect HFTs is their very high order-to-trade ratio; on the contrary, algorithmic trading, as we will see in more detail in section 1.6, can implement long term strategy with the aim to meet or beat a particular market benchmark, such as TWAP or VWAP algorithms, in order to minimize block orders' market impact. Differently from algorithmic trading, HFT trading strategies quickly liquidate their positions during the day and end the trading day with a position as flat as possible; finally, giving the short hold position of HFTs, they trade only very high liquid instruments; the following table summarize the characteristics of HFTs that are not in common with algorithmic trading:

Specific characteristics of HFT
1) Very high number of orders
2) Rapid order cancellation
3) Market making strategy
4) No significant position at the end of the day
5) Very short holding period
6) Very low margin per trade
7) Low-latency requirements
8) Use of co-location/proximity services and third parties hosting services
9) Trade in high liquid instruments

[63] SEC Concept Release on Equity Market Structure (2010)

Table 3: Specific characteristics of HFT

Source: Gomber P., Arndt B., Lutat T., Uhle T, 2011

From the above characteristics, literature on HFT has used different methods to identify high frequency traders ("HFTs"), but the two most important methods are the direct and the indirect approach.

1.5.1.1 Direct approach

The direct approach for HFT identification relies on the identification of market participants based on their primary business and the use of services to minimise latency. The primary business of a firm and whether it uses direct market feeds or co-location services can be provided by the trading venue where it operates: this criterion identifies pure HFT firms, flagged as HFT firms by the trading venue. Since HFT activity can be carried out also by firms that are not identified as HFTs, for instance investment banks, or by HFTs that rout their orders to the trading venue through non-HFT members using direct electronic access, relying on this approach can lead to an underestimation of HFT activity.[64] On the other hand, an overestimation of HFT activity can arise from the evidence that not all the activity carried out by flagged HFT firms is in fact HFT. However, since firms with HFT as primary business in all likelihood predominantly use HFT strategies, it is likely that the underestimation element will be dominant[65].

Also the second direct method to identify HFT leads to an inflation of HFT statistics on HFT activity; this approach identifies HFT as those market participants that use low-latency infrastructures, as the use of co-location and proximity services or access to fast data feeds; this approach can be too encompassing because these infrastructures are used not only by HFT firms, but also by brokers that leverage low-latency services to trade on behalf of their clients[66].

[64] ESMA, 2014, "High-frequency trading activity in EU equity markets", Economic Report, Number 1.
[65] Id.
[66] Id.

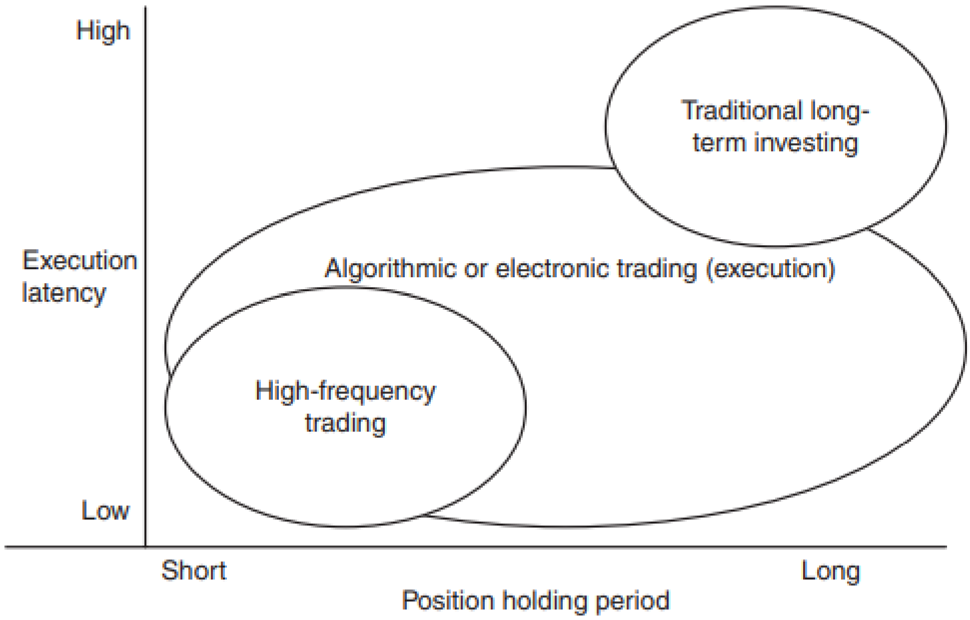


Figure 9:High frequency trading versus algorithmic trading and traditional long term investing
Sorce: Aldridge (2010)

1.5.1.2 Indirect approach

The indirect approach relies on the trading and quoting patterns of traders. Trading patterns rely for instance on the traders' inventory management; from the SEC definition of HFTs, their trading pattern must be characterized by low intraday and overnight inventory since they trade at high speed using low-latency infrastructures and close their end-day inventory as flat as possible[67]. Figure 10 illustrates an example based on mock up data on three different inventory managements; member 2 is a net seller of the stock: the negative inventory during the day and the negative end-day position means that the trader's strategy implies more short positions than long positions in the stock, an overall inventory position that he also carries out overnight. Both the other two members manage their position in order to have a flat position at the end of the day, but while member 1 has a net short position throughout the day, member 3 manages its inventory position to maintain also during the day a flat position[68]. In conclusion, with regard to traders' trading pattern, only member 3 can be identified as an HFT because it fulfills all the aforementioned characteristics: it closes the day with a flat trading position and during the day trade at high speed, trading an high volume of stocks with the aim to quickly exit the position

[67] Id. 62.
[68] Id. 63.

to minimize intraday inventory risk. However, identification based on intraday inventory management will tend to identify HFT market making strategies and may not identify other aggressive HFT strategies.

An alternative method to identify HFTs is based on the lifetime of limit orders, the time orders last in the limit order book before being cancelled or modified. Through the use of co-location and direct market feeds, HFTs are able to quickly react to changing market conditions, continuously updating or cancelling their quotes when the arrival of new informations make their quotes stale and subject to adverse selection risk from trading with informed traders. Figure 11 shows statistics from the ESMA report dataset for three categories of traders: investment banks, HFTs and other firms. The dataset encompasses 100 stocks traded on 12 European trading venues on September 2012.[69] Firms identified as HFT under the direct approach appear to send orders with shorter lifetime (40% less than 0.2 seconds), compared to Investment banks (40% less than 5 seconds) and other firms (40% less than 3 seconds). Limit orders submitted by HFTs are usually referred as "fleeting orders", a concept originated from Hasbrouck and Saar (2009), which indicates that they are added and removed from the order book within a timeframe in the order of milliseconds, making difficult for slow traders to fill them before they are cancelled or modified.[70]

The study of the trader's message traffic is another indirect method to identify HFT activity from academics, industry bodies, trading venues and regulators; under this approach, the most used proxies for HFT activity are the number of messages (submissions, cancellations and updates) per $100 of trading volume or the order-to-trade ratio (OTR), that is the ratio between the total number of orders posted in the limit order book by a trader and the number of orders actually executed; clearly, the higher is the OTR, the higher is the number of orders posted but not executed, which means that the resting orders are frequently cancelled and updated. The continuous update and cancellation of orders by HFTs characterizes their quoting patterns by an high order-to-trade ratio. With the introduction of the High-Frequency Trading act in May 2013, Germany has become the first country that regulates securities trading firms based on their trading infrastructures and order book activity characteristics. Alongside the use of low-latency infrastructures and the limited human intervention in the trading system, the third requirement a trader must fulfill to be classified as an HFT is the high intraday trading message

[69] Id.

[70] Hasbrouck, J., and G. Saar, 2009. Technology and Liquidity Provision: The Blurring of Traditional Definitions. Journal of Financial Markets, vol. 12 (2), 143-172.

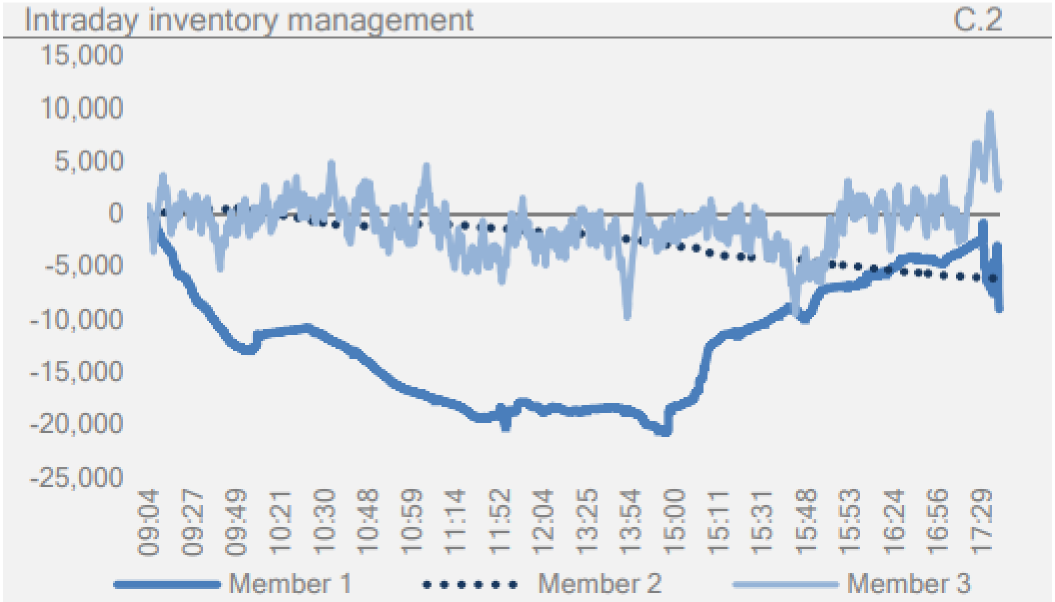


Figure 10: Net inventory of firms, based on mock up data

Source: Esma

Figure 11: Percentiles of the lifetime of cancelled or modified orders, in second

Source: Esma

volume caused by submission, quotes or cancellations[71]. In particular, the HFT German Act identifies firms as HFT if their intraday message rate exceeds 75.000 messages per trading day on an annual average.[72]

[71] German High Frequency Act, 2013; Martin Haferkorn, Kai Zimmermann, 2014, The German High-Frequency Trading Act: Implications for Market Quality, Goethe University Frankfurt.
[72] Id.

As for intraday management approach, relying only on message traffic proxy such as order-to-trade ratio may lead to biases in the results because it only considers HFTs engaged in market making strategy, where they regularly update their bid and ask quotes, failing to capture other HFT strategies such as statistical arbitrage, where traders use low-latency infrastructures even if their order-to-trade ratio is low[73]. Also algorithms used by firms for agency trading on behalf of institutional investors may result in high OTR and therefore be mislabelled HFT. Another issue linked to this method lies on the fact that an high order to trade ratio characterizes firms that operate in illiquid markets: a firm that is not an HFT can have an high order to trade ratio because trading illiquid stocks results in few trade executions despite many sent orders[74]. This result imply that HFT activity is higher for less liquid stocks than for liquid stocks, which is not in line with existing empirical evidence, which shows how HFTs are more active in large capitalization stocks than in mid and small capitalization stocks.[75] Furthermore, this conclusion is also counterintuitive: HFTs by definition trade at high speed, consequently HFTs will tend to trade in liquid markets, where it is easier to quickly liquidate an established position with no significant price effect.

Finally, the OTR does not take into account the trading speed, then this method can lead to identify as HFTs traders that have an high OTR even if they do not trade at high speed; for instance, an algorithm that updates orders every 10 minutes could have a high OTR even though it is not implementing any HFT strategy[76].

Finally, the last indirect approach method is based on the HFT trading strategy adopted by the traders; Hagströmer and Nordén (2013) distinguished HFT activity in two main groups: the market making activity and the opportunistic activity, such as statistical arbitrage or directional trading[77]. The authors use data from NASDAQ-OMX Stockholm exchange; since HFTs engaged in market making compete with other market makers to quote at the best bid and offer (BBO) prices, HFTs that quote more than 20% of BBO on a daily average across the sample of stocks, are classified as HFTs market makers, all the others are classified as opportunistic HFTs.[78]

---

[73] Id. 63.

[74] Id.

[75] See i.e. Zhang, Sarah and Ryan Riordan, 2011, Technology and market quality: the case of high frequency trading, ECIS 2011 Proceedings, Paper 95.: in their study on the information effect of HFTs they infer that HFTs bring information into the market only for high market capitalization stocks, but this effect results to be inconclusive for mid cap stocks and significantly smaller for small cap stocks.

[76] Id. 29.

[77] Hagströmer, B., L. Nordén, and D. Zhang, 2014. How Aggressive Are High-Frequency Traders?, The Financial Review, 49(2), 395–419

[78] Id.

1.6 The evolution of algorithmic trading

1.6.1 First generation algorithms: impact-driven algorithms

At the beginning of electronic trading era, the first generation of algorithmic trading was an evolution of order slicing strategy; this strategy is implemented by large institutional investors to lower their trading costs, which consists in discriminating among traders who are most willing to trade and those who are willing to trade only at inferior prices. Large traders break their orders into several pieces to trade one at a time; the first pieces trade at the best prices initially available in the order book, then the remaining pieces trade at progressively inferior prices as the traders exhaust the available liquidity and as the market discover the true order size[79]. Finally, the average at which the parent order is executed slicing it in small child orders is better than the price the large trader would have obtained if he had executed the entire order at once, given the price impact it would have generated.
The first generation of trading algorithms is called impact driven algorithms: they try to minimise the overall market impact slicing the parent order into small child orders focusing on specific benchmark prices such as time weighted average price (TWAP) or volume weighted average price (VWAP).
The average price based algorithms, TWAP and VWAP algorithms, have the object to minimize impact cost, even if their main focus is their respective benchmark; they are predominantly schedule based algorithms, and so they usually track statically created trajectories with little or no sensitivity to other market variables such as price or volume[80].

1.6.1.1 Time Weighted Average Price Algorithm

The Time Weighted Average Price benchmark is an average price which reflects how the asset's market price has changed over time; trading algorithms that are based on this benchmark usually try to meet it by slicing the parent order in several equal child orders, following a uniform time based schedule[81]. For example, the algorithm could be set to sell 20000 shares within two hours in blocks of 5000 shares, resulting in 4 sell orders for 5000 shares which are sent to the market every half an hour. Clearly, the trading pattern of such strategy is extremely uniform and independent of both market volume and price.
Trading in such a predictable way can lead to considerable signalling risk: the only thing the

[79] Harris L. ,2003, Trading and Exchanges: Market microstructure for practitioners, Oxford University Press
[80] Johnson B. ,2010, Algorithmic Trading and DMA: An Introduction to Direct Access Trading Strategies, 4Myeloma Press, London
[81] Id.

other market participants do not know is the total size of the market; furthermore, since this strategy does not consider other variables other than the TWAP, it can suffer poor execution quality when prices become unfavourable or the available liquidity sudden drops[82]. To enable traders to trade with a less obvious trading patter, traders can adopt a more flexible trading approach, randomizing the orders' size and frequency, constantly comparing order execution progress with respect to the ideal target quantity from the linear completion profile of the standard TWAP algorithm.

1.6.1.2 Volume Weighted Average Price Algorithm

As for TWAP algorithm, VWAP algorithm try to meet a benchmark, the volume weighted average price, which is equal to the ratio of the total traded value to the total traded quantity; given n trades in a day, each with a specific price p_n and size v_n, we can express the daily VWAP as[83]:

$$VWAP = \sum_n \frac{v_n p_n}{\sum_n v_n}$$

This benchmark is based on the daily trading volume, but traders do not know a priori the traded volume of the ongoing trading day; traders implementing VWAP compute the order size for each interval of the day during which they want to send the orders using the historical volume profile, the averages of historical traded volume for these time intervals during the day. If the day is divided into j periods, VWAP will be equal to[84]:

$$VWAP = \sum_j u_j \bar{p}_j$$

Where u_j is the percentage of daily volume traded and $\bar{p_j}$ is the average price in each period. Robert Kissel and Morton Glantz (2003) show that the optimal trading schedule to meet the VWAP benchmark may be based on this percentage. Thus, the target quantity x_j for each period j is[85]:

$$x_j = u_j X$$

[82] Id.
[83] Id.
[84] Id.
[85] Kissel and Glantz, 2003, "Optimal Trading Strategies: Quantitative approaches for managing market impact and trading risk", publisher American Management Association.

Where X is the total size of the order. Clearly, the use of historical data to predetermine the trading pattern implies persistency of trading volume, which means that trading volume tends to follow a similar pattern to its historical profile, and this assumption can be accepted if the historical profile is based on sufficient data, that is a reasonable assumption for many liquid stocks.

1.6.1.3 Percent of Volume Algorithm

Percent of Volume algorithm, also called volume inline, participation, target volume or follow algorithm, is set to participate in the market up to a predefined trading volume; such an algorithm could for example try to participate by trading 5% of the volume in the target instrument until it has built or liquidated a target position. Since this algorithm targets traded volume, it reflects the current market volume in its orders[86]. Unlike algorithms, such as TWAP and VWAP, where the trading schedule is deterministically determined, for POV algorithms the trading schedule is dynamically determined: the algorithm participates in the market at a given rate, in proportion with the market volume[87]: as a result, the trader no longer has any certainty about the trade completion within a specific time period because trade size is dependent on the market value of each considered time interval.

1.6.2 Second generation algorithms: cost -driven algorithms

The second generation algorithms takes the action from the transaction cost analysis (TCA); this analysis sheds light on all the components of the trading costs. The trading cost is not only determined by the price impact of the order, that was the main focus of first generation algorithms such as TWAP and VWAP algorithms, but other factors such as timing risk and opportunity cost can actually outweigh the market impact. In their attempt to reduce the market impact of large orders, impact driven algorithms slice large orders in child orders throughout the day, without taking into consideration market variables such as the price; even if splitting an orders over a long time period reduces the market impact of a parent order, on the other hand the trader is exposed to an higher timing risk, in particular when the asset is characterized by high volatility[88]: there is a trade-off between market impact and timing risk. The timing risk is the risk to buy at an excessively high price or selling at an excessively low price given

[86] Gomber P., Arndt B., Lutat T., Uhle T., 2011, High-Frequency Trading, Goethe Universitat Frankfurt Am Main
[87] Id. 79.
[88] Id.

unfavourable price movements which harms the value of investor's portfolio. Since time has opposite effect on price impact and timing risk, as shown in Figure 4, the second generation algorithms tries to tackle what Robert Kissel and Morton Glantz (2003) termed the trader's dilemma: trading too fast brings high market impact costs, whilst trading too slowly exposes traders to considerable timing risk.[89] Cost driven algorithm have to strike a balance between market impact and exposure to timing risk.

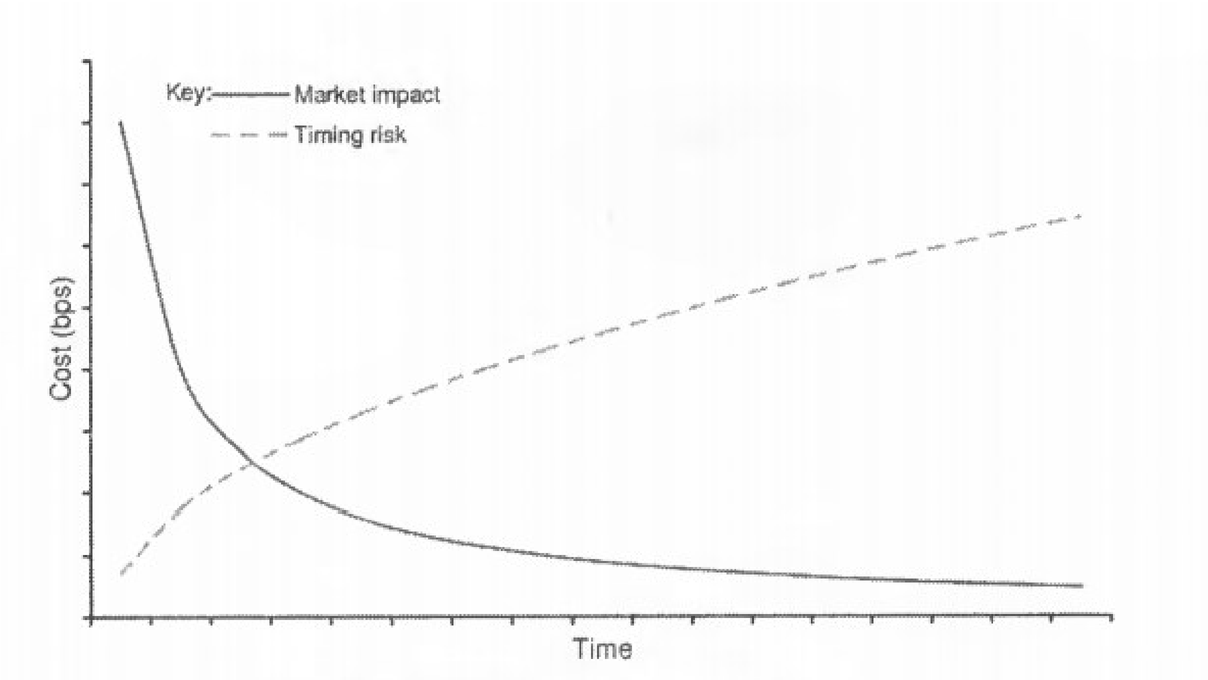


Figure 12: Market impact vs. timing risk
Source: Johnson B. (2010)

1.6.2.1 Implementation shortfall algorithm

Whereas for small trades transaction costs is usually measured by the bid-ask spread, for large orders this measure is better represented by implementation shortfall (IS), that is the difference between the price at which the order is executed and the average execution price that is actually achieved; the IS algorithm try to achieve an average execution price that minimize this difference

IS algorithm works as POV or VWAP algorithms, but differently from the impact-driven algorithms IS algorithm defines an optimal trading horizon to solve the trade-off between the aforementioned two components of the overall trading cost; this process have to consider

[89] Id. 84.

trader's and asset's characteristics, in particular the risk aversion of the former and the volatility and liquidity of the latter. Risk aversion directly affects the aggressiveness of a trading strategy: an high level indicates an higher investor's level of urgency, which means that timing risk is not acceptable and so the strategy should be more aggressive to try to complete the order execution faster, even if this implies an higher expected cost due to market impact .[90] The asset liquidity affects the bid-ask spread: the higher the liquidity of the asset the lower is the bid-ask spread charged by dealers since they can easily liquidate their positions. An higher asset liquidity implies the presence of many trading counterparties in the market and so a lower asset volatility, which reduces the transaction costs of executing several child orders over time, with low risk of price jumps given a large depth availability. From these considerations, we can conclude that a long optimal trading horizon will be suitable for liquid and low volatile assets and for traders with low risk aversion, with the opposite holding for short optimal trading horizon.

1.6.3 Third generation algorithms: Opportunistic algorithms

Opportunistic algorithms are designed to take advantage of favourable market conditions, whether this is based on price liquidity and other factors. They implement price or liquidity sensitive strategy, which enables them to adjust their trading style based on whether the current market price or liquidity is favourable or not [91].

1.6.3.1 Price inline algorithm, adaptive shortfall algorithm and liquidity driven algorithm

Price inline algorithms adapt its trading to the market price in a similar way to how POV algorithms adapt its trading to market volume; if we consider a large buy order to be executed, after having determined a benchmark price, the trading algorithm will trade more aggressively when the market price is below that benchmark, while it will do the opposite when the market price is above the benchmark, the opposite reasoning if it have to execute a large sell order. This strategy is also called aggressive in the money (AIM), from the option nomenclature term moneyness: as for an in the money option the holder of a call option will exercise the option if the strike price is below the market price, in the same way the Price inline algorithm will buy more aggressively when the market price is below a certain benchmark.[92]

[90] Id. 80.
[91] Id.
[92] Id

As price inline algorithm, also liquidity driven algorithm react to changing market conditions, where the benchmark is no longer a threshold price but a liquidity measure; the liquidity of an instrument refers to how quickly a given amount of that instrument at a given price can be traded: a market is liquid when traders can trade without significant adverse effect on price. Markets with many standing limit orders and small bid-ask spread are usually quite liquid[93]. The benchmark used by liquidity driven strategy will be a market depth metric, which reflects the available number of orders at a given price and the algorithm will follow a more aggressive strategy when this metric will be favourable[94]: for instance, a buy order will be more aggressive with plenty of market depth and low prices.

Finally, the adaptive shortfall algorithm is an enhanced version of the traditional cost driven IS algorithm: after the definition of the optimal trading horizon, the algorithm will follow an AIM strategy when prices will be more favourable for minimizing the implementation shortfall.

1.6.4 Fourth generation algorithms: Newsreader algorithms

News traders are a particular kind of informed traders that try to predict how instrument's price will change not on the basis of fundamental analysis, but on incoming informations regarding the asset. With the rise of automated trading, traders progressively increase the use of sophisticated algorithms to collect amount of information as large as possible from all the available information sources; newsreader algorithms implement statistical methods to discern among all the information collected the material information, that are informations that significantly affect instrument values. News traders' success depend not only on the capacity to discover material informations, but also on the speed at which they are able to react to these informations; the necessity of high speed determine the rise of HFTs, with an ever increasing use of ultra-low-latency infrastructure to process market data at the high possible speed to avoid to trade on stale prices.

93 Id.
94 Id.

2 HIGH FREQUENCY TRADING STRATEGIES

Many of the trading strategies implemented by HFTs are not new to financial markets, but the technological innovations brought by these traders in the trading industry computer empowered these strategies, changing the way they were implemented. The common element of the wider range of HFT strategies is the leverage of supercomputers that enable firms to execute trades at extremely low-latency, within milliseconds or microseconds.

HFT strategies can be grouped into six main categories: market making strategies, arbitrage strategies, structural strategies, directional strategies, ghost liquidity strategies and news trading; the first category includes electronic market making/spread capturing trading and passive rebate arbitrage; arbitrage strategies encompass statistical arbitrage and cross market arbitrage; structural strategies include latency arbitrage/slow market arbitrage and flash trading; directional strategies include ignition momentum strategies and liquidity detection strategies; ghost liquidity strategies include smoking, spoofing, layering and quote stuffing.

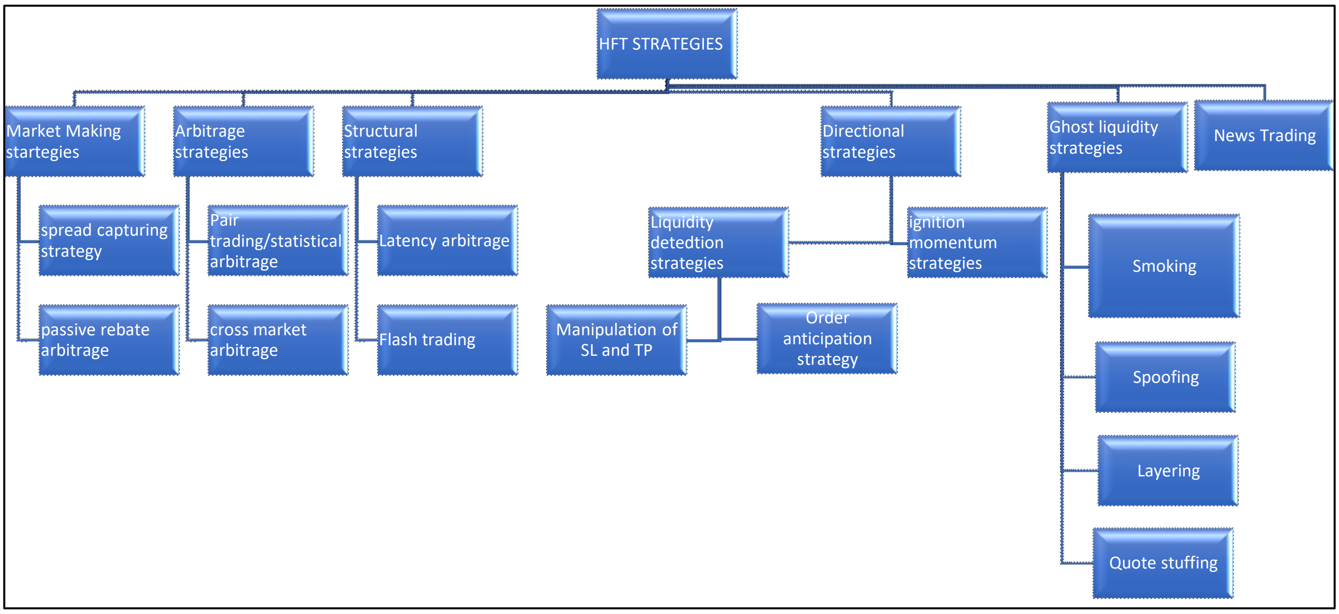


Figure 13: HFT strategies

2.1 Market making strategies

Market making is the activity of continuously quote two-sided markets, posting limit orders on both the buy and sell side of the limit order book; through this activity, market makers supply liquidity in the form of immediacy to traders who demand liquidity. Immediacy is one of the several dimensions of liquidity which refers to how quickly trades of a given size can be arranged at a given cost: the higher the liquidity of a market, the quicker is the execution of an order given its size and price. [95] When traders want to quickly trade they submit market orders that are executed at the available BBO or BBA at the time of execution, respectively for a buy and sell order, posted by the market maker on the limit order book.

Before the rise of ECNs, market makers were designated exchange's members, such as NYSE specialists and Nasdaq dealers; by posting limit orders, they allow other traders to trade when they want to trade: market makers profit when they buy from impatient sellers at lower prices than they sell at impatient buyers[96]. The difference between the price at which the dealer buys from the sellers, the bid, and the price at which he sells to the buyer, the ask, is called the bid-ask spread and it is the compensation they receive for offering immediacy.

ECNs allow investors to have direct electronic access to marketplaces, enabling them to directly trade one another, even acting as market makers, replacing members' liquidity supply. HFTs implement this strategy for two purposes: they supply liquidity to earn the bid-ask spread, that is the stricto sensu electronic market making, and to leverage maker-taker fee structures adopted by some trading platforms.

2.1.1 Spread capturing strategy

Under their activity to continuously posting limit orders, market makers face adverse selection risk, the risk to trade with informed traders, that is the risk that prices will move against their position after having traded with this kind of traders: informed traders trade when prices deviate from their fundamental values, so that they buy an asset when it is undervalued and sell it when it is overvalued. Thus, market makers have incentives to make sure that their limit orders reflect as much current information as possible as quickly as possible, in order to avoid losses trading with informed traders[97]. HFTs progressively replaced designated market makers: using low-latency infrastructures, they are able to quickly react to changing market conditions and to collect as large as possible volume of market data, continuously updating their limit orders and drastically reducing their losses from trading in the wrong side of the market; consequently, since market makers quote wider bid-ask spread the higher is the adverse selection risk, the reduction in the adverse selection risk results in a narrower spread than that quoted by non-HFT registered market makers.

Since they are not registered as exchange's market makers, HFTs are informal liquidity providers and consequently they are not subject to the designated market makers' affirmative

[95] Id. 79.
[96] Id.
[97] Charles M. Jones, 2013, "What do we know about high frequency trading?", Columbia Business School.

obligations. Affirmative obligations oblige registered market makers to trade when no other traders are willing to trade, in order to prevent disorderly markets, avoiding price jumps and excessive market volatility: for this reason registered market makers are usually called traders of last resort.[98] Market makers have to continuously quote two sided-markets in the security or securities for which they are registered on a trading venue and they have to intervene to ensure price continuity, preventing large price reversals; price reversals occur when prices rise and then fall or fall and then rise, generating excess volatility. Price reversals are due to lack of liquidity and designated market makers work to ensure that prices move smoothly, without jumping too much: price continuity helps ensuring public traders that brokers fill their agency orders fairly, not at irrational prices (stub quotes)[99].

2.1.2 Passive Rebate Arbitrage

Exchanges and ATSs compete to attract order flows not only by offering to investors highly technological services that improve the execution of the orders, such as co-proximity services or direct data feeds, but also adopting fee structures that subsidize the provision of liquidity in the market. One of the key factors that investors examine to evaluate the quality of a trading venue is the market depth at different price levels; the higher the number of limit orders at a given bid or ask price, the higher is the capacity of the market to absorb large orders without significant price impact. Thus, the depth of the market represents a competitive advantage for trading venues and consequently they began to offer fee structures to attract liquidity providers. The fee structure that trading venues use to attract market makers is the maker-taker model: adopted for the first time by Island ECN in 1997, it is an asymmetric pricing model that charges an access fee on liquidity demanders (takers) for executing their market orders that fill against (take) standing limit orders and provides a rebate to liquidity providers for their executed standing orders that make markets; the difference between the access fee and the liquidity rebate is the net fee that maker-taker exchanges earn for arranging trades[100].

HFTs are incentivized to provide liquidity to trading venues with maker-taker fee structure with the aim to capture as much of the liquidity rebates as possible; in particular, since liquidity rebates are usually fraction of a cent per share, they can leverage their technological advantage to trade large volume of orders at ultra-fast speed in order to earn a significant margin. For example, the NYSE charges $0.0024 per share for aggressive marketable orders that execute against passive orders on the trading floor. However, this fee will be $0.00275 for non-floor

---

[98] Id. 79.
[99] Id.
[100] J. Angel, L. Harris e C.S. Spatt, Equity Trading in the 21st Century, USC Marshall School of Business, 2010.

transactions and \$0.0030 if you are a "designated market maker." These fees also depend on traders' average trading volume.[101]

Furthermore, with maker-taker fees HFTs can trade more aggressively, narrowing the bid-ask spread; for example, if S is half of the bid-ask spread and P is the bid-ask spread midpoint, in the absence of a liquidity rebate HFTs will post an ask equal to $P + S$ and a bid equal to $P - S$. If the exchange offers a liquidity rebate equal to R ($R < S$), the ask will be $P + S - R$; this is because when HFTs receive a liquidity rebate for every limit order executed, from an economic point of view receiving $P + S - R$ is equivalent to receive $P + S$ with no rebates. Following the same reasoning the bid will be equal to $P + S + R$. Thus, the presence of a maker-taker fee structure incentivizes HFT market makers to adjust limit buy order up by R and limit sell order down by R, narrowing the bid-ask spread[102].

2.2 Arbitrage strategies

Arbitrage strategies are trading strategies that exploit price discrepancies between correlated instruments; since these market inefficiencies last for very short periods, HFTs' ultra-fast speed trading can successfully implement arbitrage strategies, making a profit after the convergence to instruments' normal relationship.

2.2.1 Pair trading

Pair trading is an arbitrage strategy that try to exploit mispricing between a pair of similar, highly correlated instruments; when the basis, the difference or ratio between the prices of the two instruments, is different than the fair value of the basis, usually the historical spread or ratio between the two instruments' prices, pair traders buy the cheaper instrument and short the more expensive one: arbitrageurs profit if either the price of the cheaper one goes up and the price of the more expensive one goes down (purchases are undervalued and the sales are overvalued) or if the cheaper one appreciates faster than the more expensive one or if the more expensive one depreciates faster than the cheaper one (purchases are undervalued relative to sales).[103] When the basis converges to its fair value, pair traders lock in profits by trading the reverse pair to flatten their position. Usually, pair traders use the historical mean of the spread or price ratio as

---

[101] Ilan Guedj, Zhong Zhang, Maker-Taker Fees in A Fragmented Equity Market. 2019.

[102] Fox, Glosten, Gabriel, Rauterberg, 2015: The New Stock Market: Sense and Nonsense, Duke Law Journal, Vol. 65, No. 2, 191-277.

[103] Id. 79; similar instruments' value depends on common factors, which may include macroeconomic variables (interest rates, national income, unemployment, expected inflation), industry variables (sales, wages, prices, product innovations, competitive conditions), physical variables (weather, agricultural pests, solar activity) and political variables (legislative, executive, judicial, military interventions).

fair value of the basis: with the assumption that market efficiency causes mean reverting asset returns, basis tend to be mean reverting in the long run. A trade entry signal can occur when the spread exceeds a predetermined band around the mean level, for example two standard deviations as in figure 14, after which arbitrageurs expect that the spread begins to converge to the mean. The correct identification of the entry signal allows arbitrageurs to avoid losses due to the further widen of the spread after having established the position[104]. Arbitrageurs may incorrectly identify mispriced correlated instruments when the relative change of one instrument is due to specific factors of this instrument, or when the fundamental relationship between them is changed.

Pair trading is a market neutral strategy because pairs traders simultaneously buy and short positive correlated instruments, protecting themselves from price changes due to common factors that influence in the same way both the instruments: any losses on the short position due to an increase in the price of the instrument is hedged or offset by gains on the long position, and vice versa.

Correlated instruments are inconsistently priced when the price of one instrument is slower than the other in adjusting to common factors changes, or when uninformed traders' trades move prices away from their fundamental value; by exploiting these price inconsistencies arbitrageurs enforce the law of one price, making prices more informative: when they buy the undervalued instrument they push up its price toward its fundamental value, while when they sell the overvalued one they lower their price to their fundamental value[105].

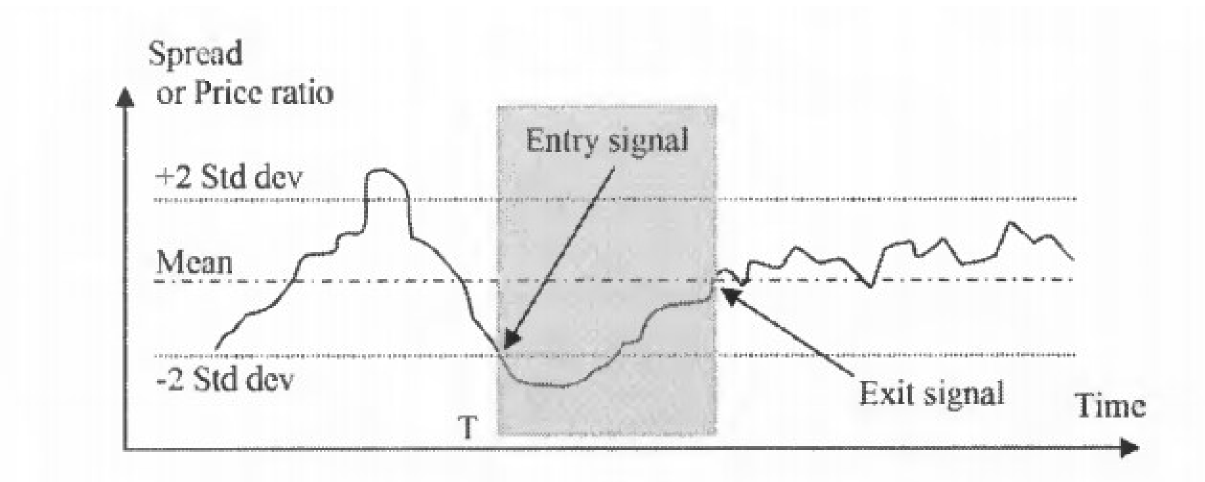


Figure 14: Pair trading strategy

Source: Johnson B. (2010)

[104] Id. 80.
[105] Id. 79.

Pair arbitrage trading can involve shares of companies operating in the same industry, for example Ford common stock and GM common stock are affected by automotive industry fundamentals and other macroeconomic factors such as interest rates and labor conditions; index arbitrage involves a stock index, for example the S&P 500 and a portfolio of individual securities composing the index, weighted according to their weights within the index. Under the law of one price stock index price should be equal to the price of the index-mimicking portfolio[106]. Pair arbitrage can be exploited between a stock index and an ETF (exchange traded fund), an investment fund that tracks the performance of an underlying stock index; a well known example is the SPDR S&P 500 ETF (SPY) which tracks the S&P 500 index. Pair arbitrage can also be applied to pairs consisting of a security and its derivative; for example the derivative of choice is often a futures contract since futures prices are linear functions of the underlying asset: $F_t = S_t\, exp[r_t(T - t)]$, where F_t is the price of a futures contract at time t, S_t is the price of the underlying asset at time t, T is the time the futures contract expires, and $r_t(T - t)$ is the interest rate from time t to T[107]. When the previous condition is violated a pair arbitrage can be exploited; for example, if the LHS is lower than the RHS an arbitrageur can enter into a long position in the futures contract with delivery price F_t, short the underlying asset and invest the revenue from the short selling, S_t, in a bank account up to time T at the interest rate $r_t(T - t)$; the net cash flow of these three operations at t is null. At maturity the futures contract delivers its payoff $S_T - F_t$, the arbitrageur pays back the current market value S_T of the underlying, and he gains from the bank account $S_t(1 + r_t(T - t))$: the net cash flow at maturity T is equal to the risk-free (arbitrage) profit: $S_t(1 + r_t(T - t)) - F_t > 0$[108].

2.2.2 Statistical arbitrage

Statistical arbitrageurs use factor models to generalize the pairs trading strategy to many instruments; factor models are statistical models that represent instruments' return by a weighted sum of common factors plus an instrument specific factor. The factors used in their statistical models include macroeconomic variables such as interest rates, inflation rates, industrial production, credit spreads, stock index levels, and market volatility, that are common factors portfolio's correlated instruments' prices depend on. Statistical arbitrageurs estimate current factor values in order to determine which instrument is inconsistently priced with

[106] Irene Aldridge (2010), High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems, John Wiley & Sons, Inc.
[107] Id.
[108] Id. 49.

respect to their common factor representation. Finally, the arbitrageurs sell the instruments that are overpriced and buy those that are underpriced[109].

2.2.3 Cross markets arbitrage

The fragmentation of financial market in many trading venues has increased cross markets arbitrage opportunities: since the same instrument can be quoted in different trading centers, cross market arbitrageurs can profit when they are quoted at different prices in different venues by buying the cheaper and selling the more expensive; after the convergence to the same price they close their positions making a risk-free profit. Usually, price discrepancies between quotations of the same instrument in different trading venues last for very brief period (fraction of seconds); HFTs leverage co-location services and direct market feeds located in multiple trading venues to quickly scan and spot cross market arbitrage opportunities, implementing the strategy before others can correct the inefficiencies. Thus, it is clear the winner-take all nature of arbitrage-oriented HFT: HFT arbitrageurs that are consistently faster than any other market participant will be able to quickly buy up undervalued instrument and sell overvalued ones, bringing their prices to the unique fundamental value, leaving slightly slower traders with no arbitrage opportunities to exploit[110].

2.3 Structural Strategies

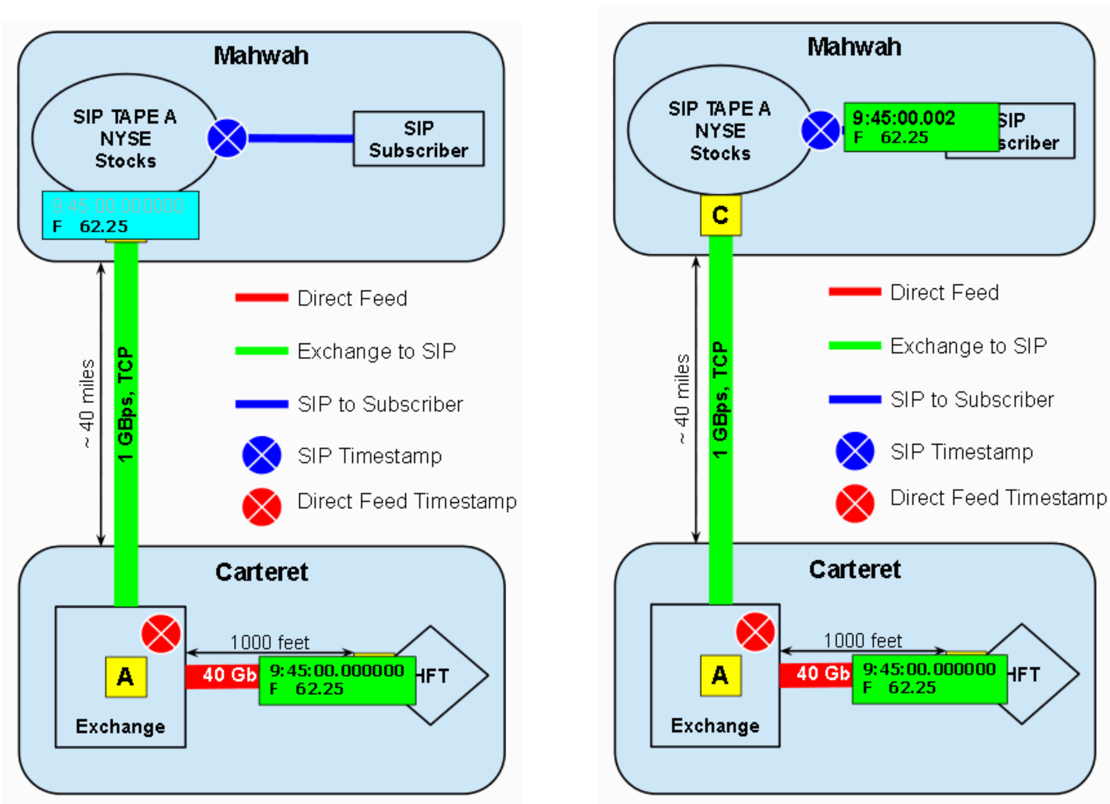
HFTs leverage low-latency infrastructures to exploit structural vulnerabilities in the stock market; obtaining the fastest delivery of market data through co-location arrangements and individual trading center data feeds HFTs implement two structural strategies: latency arbitrage and flash trading.

2.3.1 Latency arbitrage

Regulation NMS states that "any national securities exchange, national securities association, broker, or dealer that distributes information with respect to quotations for or transactions in an NMS stock to a SIP, broker, dealer or other persons shall do so on terms that are not unreasonably discriminatory"[111]. The SEC's interpretation of the provision has been that "distributed data could not be made available on a more timely basis to private clients than

[109] Id.
[110] Id. 97.
[111] 17 C.F.R. § 242.603(a)(2).

core data is made available to the SIP"[112]. Thus, Reg NMS prohibits SRO and broker-dealers to transmit data to a private market participant before they have transmitted them to a Network Processor. The SEC's interpretation of "unreasonably discriminatory" lies on when the trading venue sends the data, not on when traders actually receive them. Market participants that leverage co-location or direct market feeds receive market data slightly before those which use the SIP, even if trading centers send market data to direct data feeds and the SIP simultaneously. The lags of SIP to transmit data to SIP subscribers with respect to direct data feeds data transmission to private clients is due to the large volume of market data from all the trading venues using the consolidated stock market data stream that the SIP has to process to provide investors a unified view of U.S. stock market prices and volumes, in particular in order to provide the NBBO. Consequently, even if market data reach SIP and direct data feeds with the same timestamp, while SIP consolidates all the market data from different exchanges and alternative trading system, market data feeds will transmit the updated data to HFTs a few milliseconds before. As shown in Figure 15, even if data on stock "F" has reached SIP and the direct data feed at the same time, 9:45 a.m, because of the consolidation process of the SIP, those data will reach the SIP subscriber with a 2 milliseconds delay.



[112] Regulation NMS, 70 Fed. Reg. 37,496, 37,567 & 37,569 (June 29, 2005).

Figure 15: Direct vs SIP Data Feed
Source: Nanex

The resulting time difference offers HFTs the opportunity to exploit latency arbitrage, also known as slow-market arbitrage. Latency arbitrage occurs when an HFT is posting the NBO or NBB on a trading venue and later on a trader posts a quote that improve the HFT's NBO or NBB on a different exchange. From the order protection rule[113], an incoming order must be executed at the NBO or NBB, respectively for a buy order and a sell order; through the use of co-location services, HFT is able to learn of the better quote before it is reported by the national market system. In the short time before the SIP reports the better quote, if the incoming order arrives at the exchange where the HFT is operating, it will transact against the now stale quote of the HFT; consequently, the HFT will close their position transacting against the better quote on the other exchange[114].

The following example will clarify the mechanism of the latency arbitrage; suppose that Virtu Financial, one of the major US HFT firms, is quoting the NBB of Amazon at the NYSE equal to $121,15; subsequently a trader improves HFT NBB posting a limit buy order for Amazon stock equal to $121,16 on BATS. Through co-location services offered by BATS, HFT computers located close to the trading venue's servers send an ultra-fast message to Virtu Financial informing it of the new NBB. As long as the better quote is not reported by the national market system, the HFT's buy order is still the NBB; thus, an incoming market sell order transacts against it. Virtu Financial co-location facility at NYSE then send an ultra-fast sell order to BATS that transacts against the other trader's $121,16 bid, making a profit equal to $0,01.

Differently from statistical arbitrage, where arbitrage opportunities follow an analysis of instruments' fundamental values, latency arbitrage is only based on technological advantage in terms of trading speed; in its famous book "Flash Boys", Michael Lewis harshly criticized latency arbitrage, defined as a predatory trading strategy carried out at the expense of slow traders.[115] In the above example, while the HFT is better off than the case it had not implemented the latency arbitrage, making a profit of $0,01, the incoming sell order is executed at an inferior price with respect to the price at which it would be executed if the HFT strategy was not implemented.

[113] 17 C.F.R. §242.611(a)(1).
[114] Id. 102.
[115] Michael Lewis (2014), Flash Boys, W. W. Norton & Company.

2.3.2 Flash Trading

One of the pillars of SEC's Regulation NMS is the best execution of investor's order: any order in NMS stocks must be executed at the NBBO; consequently, exchanges that cannot execute marketable orders at prices as good as or better than the NBBO must route these orders to exchanges offering it. Flash trading represents an exception to the NMS trade-through rule: as we have seen in section 1.1.2.1, SEC requires trading venues to make available to all market participants their best bid and offer in order to enable them to route their orders to the one posting the NBBO; however, Regulation NMS allow trading venues to not make their best bid and offer available to all market participants if "they are immediately executed after communication".[116]

Many electronic exchanges use flash order facilities to solicit trading interest among their local dealers when they receive marketable orders in securities for which they are not posting the best available prices.[117] Flash orders are marketable orders that an exchange grants to qualified traders (those able to trade at high-frequency) for a fee when the NBBO is not available at the exchange or when the exchange cannot fill the order at the NBBO in its entirety; during the flash period, an interval that lasts for fraction of seconds (between 30 to 150 milliseconds, depending on the exchange), exchanges give qualified traders the opportunity to execute the order at the NBBO or better price; if no qualified traders are willing to execute flash orders or they are not able to fully execute it, either the exchange will route the orders to trading venues where the NBBO is available or it cancels the order if so specified by the submitter, for example with an immediate-or-cancel or fill-or-kill instruction attached to the order.[118]

Exchanges use flash trading facilities to increase their execution rate and consequently their revenues from order flows; flash trading option also benefits the submitters of the orders: it improves their execution price avoiding to route their orders towards exchanges with maker-taker fee structures which charge access fee to liquidity takers. Furthermore, the faster execution of flash trades allows submitters to avoid the execution risk that their order will not be fully executed or it will be executed at inferior prices because the better quote was filled or cancelled during the order intermarket routing.

On the other side of flash trading, HFT market makers leverage this practice to earn the bid-ask spread and to selectively offer liquidity based on current market conditions and order sizes

---

[116] 17 CFR § 242.602(a)(1)(i)(A).

[117] Lawrence E. Harris, Ethan Namvar, 2016, The Economics of Flash Orders and Trading, Journal Of Investment Management, Vol. 14, No. 4, 1-17.

[118] Id. An immediate-or-cancel (IOC) order has to be executed immediately and fully, or as fully as possible. Non executed parts of an IOC order are deleted without entry in the order book; a fill-or-kill order has to be executed and fully or not at all. If immediate and full execution is not possible, the order is deleted without entry in the order book.

without committing to offer liquidity to any trader who submit market orders at the exchange. Dealers may are willing to trade at the NBBO or better price only with small retail traders, not with large institutional ones whose trades are large enough to move the market and thus impose losses on their inventories.[119]

There are other three reasons why HFTs may decide to become the counterparties of flash trading: first, if an HFT is long or short on the stock offered through flash facilities, it can close its position at a price before it is available to the rest of the market; second, if an HFT has a flat position on the stock, the flash order offers it an arbitrage opportunity: for the time the flash order is available, the HFT will try to establish a position at an higher or lower price than respectively the flash sell and buy order; for example if a fill-or-kill flash buy order for 1000 stocks of GE must be executed at $30,20, the HFT will try to establish a long position of the same size at a price lower than $30.20. Third, in the case that the HFT is posting limit sell orders on other exchanges, if it considers the flash buy order would cause an high demand pressure that can be transferred to other exchanges, it will cancel the pending sell orders to avoid to be short in a market with buying pressure; the opposite reasoning holds for a flash sell order[120]. In conclusion, flash trading offers to HFTs a sort of latency arbitrage: the informational advantage offered to HFTs for a fraction of second allows them to obtain a substantially risk-free gain (a true “free lunch”).

Despite the aforementioned benefits for exchanges, submitters and market makers, flash trading impairments the competition within the National Market System; it reduces the intermarket routing, thus the incentives to quote aggressively at secondary exchanges that do not receive as much order flows as larger exchanges. This leads to less liquidity in secondary exchanges, which increases bid-ask spread worsening trade execution. Furthermore, the withdrawal of market makers from secondary exchanges decreases their competition with primary exchanges, which in turn further deteriorate public welfare by charging higher fees[121].

2.4 Directional Strategies

Directional strategies are trading strategies that try to gain from anticipated price movements by previously establishing an unhedged long or short position, depending on the direction of the prices. Directional strategies encompass different trading strategies; when informed traders’ fundamental analysis discovers an asset’s price is deviating from its fundamental value,

---

[119] Id.

[120] Alfonso Puorro, 2013, High Frequency Trading: una panoramica, Questioni di Economia e Finanza n° 198, Banca d’Italia, 7.

[121] Id.

informed traders implement directional strategies: they take a long position on the asset if it results undervalued or they take a short position if on the contrary it is overvalued. Thus, if they have correctly anticipated the price movement, they close their position with a profit when the price will converge to its fundamental value.

Beside the above straightforward directional strategy, there are other two more complex categories of directional strategies: liquidity detection strategies, which include order anticipation strategy and manipulation of stop loss and take profit orders and momentum ignition strategy; the former try to anticipate price movements testing market liquidity with pinging strategies, the latter instigates other market participants to trade aggressively, causing the momentum trader's desired price movement.

2.4.1 Liquidity detection strategies

2.4.1.1 Order anticipation strategy

Order anticipation strategy involves the use of computer algorithms to identify large orders that sit in dark pools or other trading venues; while dark pools are trading venues specifically shaped for large traders, since they give them the opportunity to trade block trades in fully anonymity, large traders can also execute their orders on light trading venues, slicing the hidden parent orders in several child orders. Deploying the so called pinging strategy, HFTs repeatedly submit small-size exploratory trading orders intended to detect orders from large institutional traders; once HFTs receive a ping or a series of pings, which means that the orders have been executed, they are alerted on the existence of an hidden large buy side investor's order. Consequently, HFT will trade ahead of the large order in order to benefit from its market impact[122]. For example, if the pinging strategy reveals the presence of a large sell order on a particular stock (the exploratory small buy orders have been executed), HFTs will aggressively execute sell orders on that stock before the large sell order's market impact will significantly decrease stock's price; then, HFTs will close their short position directly trading with the large seller, buying from him at a price lower than the average price paid for the previously established short position. The opposite strategy will be implemented in the case of the detection of a large buy order.

This form of liquidity detection strategy is also referred as electronic front running[123] to distinguish it from the illegal front running of block transactions, that is defined by the Financial

[122] Gary Shorter, Rena S. Miller, 2014, High-Frequency Trading: Background, Concerns, and Regulatory Developments, Congressional Research Service, 23.
[123] Id.104.

Industry Regulatory Authority (FINRA), the self-regulatory organization of broker-dealers, as the illegal activity of a trader that executes buy or sell orders by virtue of material, non public informations concerning a block transaction before these informations have been made publicly available or have otherwise become obsolete or stale.[124] From FINRA definition it is clear how front running is a particular form of insider trading, which involves a trader in possession of material non-public informations about a company trading in the company' securities to make a profit or avoid a loss. However, electronic front running is not illegal: the ratio of insider trading law is a breach of a fiduciary duty, as when a broker trades ahead of its client's order making the client's order execution less profitable, or when the manager of a company trades on company's securities before an announcement by the company that will likely affect its stock's price; in electronic front running there is no breach of fiduciary duty since HFTs gather information about the flow of order by pinging different markets, which means sending multiple orders in different markets to see whether they will be filled, which can give an indication of the future direction of a stock.[125]

2.4.1.2 Manipulation of Stop Loss and Take Profit orders

Liquidity detection strategy implemented by algorithmic trading (not necessarily HFT) tests market "key" levels through the issuing of small orders, to verify the presence of stop loss or take profit orders. Market key levels refer to trading levels as support and resistance of technical analysis: support is a price level where a downtrend can be expected to pause due to a concentration of demand or buying interest; on the contrary, resistance zones are price levels where an uptrend can be expected to pause due to selling interests; another key level is also represented by price levels at which many trades have occurred during the day.[126] Algorithmic traders try to test these price levels because they are likely to serve as potential points triggering stop loss or take profit orders. Stop loss order, as the name suggests, is an order type used to limit losses given to adverse price movements: a stop instruction stops an order from executing until price reaches a stop price specified by the trader; in particular, a stop buy order closes a short position if the stock price rises to the stop limit order, while a stop sell order close a long position if the stock price falls to the stop limit price[127]. Take profit order is a limit order which specifies the price at which to close out an open position for a profit. Many

[124] FINRA 5270(a)
[125] Peter Henning, "High-Frequency Trading Falls in the Cracks of Criminal Law, New York Times, April 7, 2014, available at http://dealbook.nytimes.com/2014/04/07/high-frequency-trading-falls-in-the-cracks-of-criminal-laws/.
[126] Id.120.
[127] Id. 79.

traders use take profit orders in conjunction with stop loss orders to manage their open positions. Algo traders try to detect stop loss orders or take profit orders submitting market orders at the aforementioned market key levels in order to execute orders at prices that are expected to trigger these kind of orders; once the stop limit order or the take profit limit order is reached, the posted stop and take profit orders are activated flooding the market of liquidity; eventually, algo traders accumulate the liquidity offered at those closing prices.

Differently from the order anticipation strategy, algo traders implementing liquidity detection strategies are indifferent of whether to previously establish a long or short position: the aim of liquidity detection strategy is not to trade ahead of a large trader, but to accumulate a position on a particular financial instrument[128].

2.4.2 Ignition momentum strategy

One of the finest directional strategy is the so called ignition momentum strategy; the momentum trader (usually HFT, but not necessarily) takes an aggressive position on a financial instrument with the aim to ignite a price movement in a particular direction to instigate other market participants to trade at artificially high or low prices, respectively if the momentum trader takes a long or short position. Once the desired movement is triggered, the momentum traders close the previously established position by gaining profits.

Ignition momentum strategy is characterized by an initial stable price, then a sudden increase of volume, followed by a strong price movement either up or down and finally price reverses to the initial level. Figure 16 shows the different stages that characterize a momentum ignition strategy; in particular, it analyses a momentum ignition strategy occurred on July 13, 2012 on Daimler stock, listed on XETRA[129].

[128] Id.120.
[129] Id.

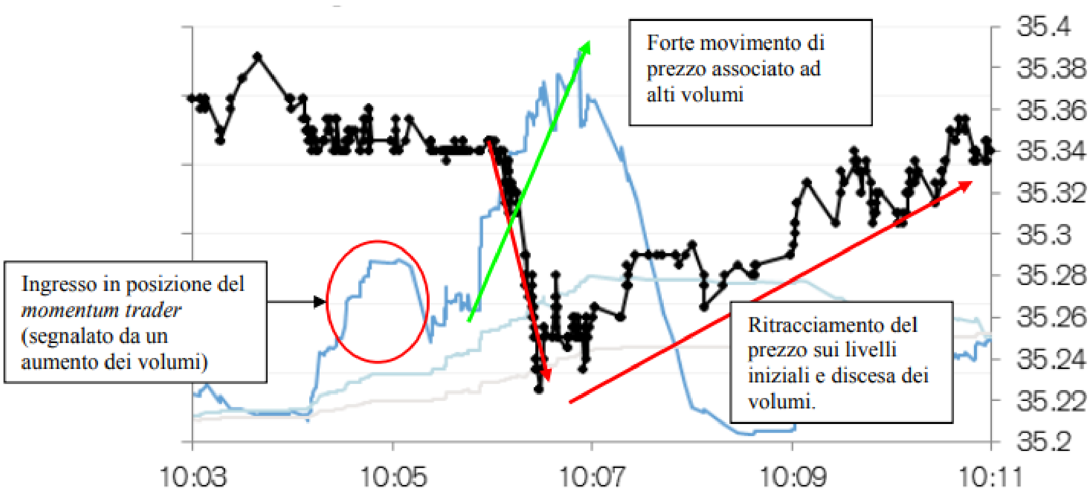


Figure 16: Ignition momentum strategy

Source: Alfonso Puorro (2013)

The initial phase of the strategy is characterized by an increase in the traded volume due to the entry in the market of the momentum trader, which starts to accumulate a position on the stock (in this case a short position); during the period of the accumulation of the position there is no significant price movements: the price of the stock remains relatively stable at $35,35 between 10:04 and 10:06. Once the storage phase is finished, momentum trader sends to the market an high volume of orders; this orders, strategically sent during a period of low volatility and low bargaining volume in order to amplify the effect of the strategy, solicit other market participants to react to the changing market conditions by submitting market orders to close their positions. The momentum trader, which have previously established a given position on the stock, fills the closing orders of the other market participants to close with a profit its strategy; in the example of Figure 16, the closing orders of the other traders are market sell orders (the stock price significantly decreases to $35,22 in less than a minute) and the momentum trader can cover its previously established short position matching these sell orders. Ultimately, as the red line on the figure shows, prices reverse to the starting level within a few minutes[130].

2.5 Ghost liquidity strategies

Ghost liquidity strategies are trading strategies that were born with the rise of HFT; they are ultra-low latency strategies implemented with the object to give a misrepresentation of the actual deep of the trading book, creating an artificial liquidity that manipulate market prices.

[130] Id.

The liquidity generated by these strategies is called ghost liquidity, or phantom liquidity: it is a liquidity provided by HFTs that is fleeting and transient due to the posting and then the almost immediate cancellation or modification of limit orders. By flooding the market with limit orders, continuously modifying or cancelling them, HFTs simulate market situations that induce other market participants to trade in an erroneous way as a reaction to false market signals. HFTs act as counterparts of these traders and after having stored the target position, they replicate the same manipulative strategy on the opposite direction, inducing other traders to close the position previously established (usually suffering a loss given the misleading signal that has led them to trade), closing their positions by gaining a profit.

Many times, ghost liquidity strategies are used to identify algorithmic traders' recurring investment patterns. For example, let assume an institutional trader (a pension fund or mutual fund) that trades following an investment scheme based on a specific algorithm; an HFT that wants to close its short position, submits many small-size sell limit orders with different limit prices cancelling them after a few millisecond; with this strategy the HFT can observe the reactions of the algo trader in order to learn its investment scheme. Once the HFT has become aware of the scheme, it begins to simulate the inputs that lead the algo trader to trade, forcing it to buy when the HFT needs liquidity to close the short position[131].

There are three main ghost liquidity strategies: smoking, layering and spoofing. Smoking strategy consists of sending tempting orders to the market in order to attract slow market participants to trade in a particular direction; however, these orders are fake orders that are quickly replaced with orders at inferior prices before the slow trader can become aware of the changed scenario.

Layering represents a more sophisticated ghost liquidity strategy; let assume that the HFT wants to buy a stock at a better price than the current best ask in the trading book; the HFT will start to send to the market limit sell orders at progressively decreasing limit prices with the aim to induce other traders to think that the market has initiated a bear phase on that stock. Once other traders' desired behaviour is triggered, the HFT use its ultra-fast trading infrastructure to cancel the submitted orders before they can be executed; meanwhile, the HFT submits limit buy orders that will be filled by the incoming sell market orders at better prices than the initial best ask. Eventually, the HFT implements the reverse strategy to close the position with a profit: it will submit limit buy orders at progressively higher limit prices to simulate the situation of a bear market; once the other traders will be influenced by the reverse spoofing strategy, the HFT will quickly cancel its orders and will post limit sell orders at the improved prices provided by the

[131] Id.

other traders, closing its position by gaining a profit.

Spoofing is a trading strategy similar to layering, but differently from the latter it uses a low number of orders; the HFT places an hidden small orders on one side of the trading book, for example a buy order, and a large light order on the other side of the market in order to induce other traders to react to the artificial bull market generated by the HFT; as for the layering strategy, the HFT leverages its ultra-fast trading system to cancel the light order before it is executed and it executes the hidden orders at the artificially deflated prices (we will discuss in more details these two strategies and their implications in section 4.8.1).[132]

2.5.1 Quote Stuffing

Quote stuffing is the HFT strategy that consists in flooding the market with an overwhelming volume of orders that are immediately cancelled; this excessive volume of orders clogs the market's data pipes with unnecessary messages that do not result in trades. The congestion of the market through the submission and cancellation of thousands of orders has the consequence to slow down the trading systems of slower traders, that differently from HFTs have not the computational power to process a large volume of data at high speed, and to slow down the overall functioning of the trading venue where HFTs operate[133]. Consequently, Quote Stuffing creates exploitable latency arbitrage opportunities to HFTs; the latency arbitrage opportunity arises from requiring other traders to process large amounts of volume, giving the HFTs submitting the orders an advantage. A large number of order submissions may also cause the exchange receiving the quotes with lags with respect to other exchanges, creating cross market arbitrage opportunities. It is also possible that large bursts of quoting activity may not be caused by manipulative trading; large spikes in quoting activity can be due to the interaction of two or more algorithm that interact one another; for example an algorithm's quote activates the input of another algorithm that reacts, causing the first algorithm to respond. If these algorithms continue to chase one another they generate a large burst of quote similar to that generated by Quote Stuffing strategy[134].

To clarify this HFT strategy we discuss the Quote Stuffing strategy carried out by Citadel Securities, LLC, one of the leading market maker in the world, on February 12, 2014. On June 16, 2014, Nasdaq posted a disciplinary action against Citadel Securities, LLC (CDRG); Nasdaq stated that during the period from March 18, 2010, to February 28, 2014, CDRG failed

---

[132] Valeria Caivano, Salvatore Ciccarelli and Giovanna Di Stefano, 2012, Il Trading ad Alta Frequenza: Caratteristiche, Effetti, Questioni di Policy, CONSOB Discussion Papers No. 5, 22.

[133] Id.

[134] Jared Egginton, Bonnie F. Van Ness, and Robert A. Van Ness, 2016, Quote Stuffing, Financial Management.

to establish, maintain and enforce a supervisory system and risk management controls reasonably designed to manage the message activity of its fully-automated electronic trading desk. Among the several instances during which CDRG' s equity market making desk failed to prevent the transmission of erroneous orders, Nasdaq analyzed the one occurred on February 13, 2014, that it concluded was caused by a glitch in CDRG's control system.

As reported by Nasdaq disciplinary action, after a code rearrangement project completed in January 2014, CDRG misconfigurated a software code that led to the de-activation of a "trashing control" system, a control system designed to cap the maximum number of orders that could be sent to the market at 200 orders per second in a given security. That day, between 13:32:53:029 and 13:33:00:998 CDRG transmitted to Nasdaq approximately 8-9 orders to buy 100 shares of Penn National Gaming, Inc ("PENN") every microsecond for a total of 65,000 orders with zero execution[135].

Differently from Nasdaq, Nanex, a Chicago-based firm that offers streaming market data services and real time analysis and visualization tools, did not think that the excessive message activity was caused by a glitch of the CDRG's threshold system, but by a Quote Stuffing strategy intentionally implemented by CDRG.

Figure 17 shows the latency arbitrage opportunities offered by the Quote Stuffing strategy of CDRG; the figure reports the quotation activity of Total View (the Nasdaq Direct Feed) and SIP within one second of the interval during which the Quote Stuffing strategy was implemented. Each pixel of the x axis is equal to one millisecond; the blue line is the number of quotations generated by the SIP every millisecond, while the red line tracks the number of messages registered by the direct feed of the Nasdaq; the green line shows the lag between the SIP and the TotalView in reporting quotes. In accordance with Reg NMS there should be not lags: market data cannot be made available to clients using direct data feed before they are made available to investor's using the SIP; however, because of the congestion generated by Quote Stuffing, there are SIP delays that can be even higher than 10 milliseconds[136]. The delay in the SIP quote reporting allows HFTs to exploit latency arbitrage opportunities and

[135] The Nasdaq Stock Market LLC, Notice of Acceptance, Waiver and Consent (AWC), June 16, 2014.
[136] NANEX, The Quote Stuffing Trading Strategy, August 15, 2014, link: http://www.nanex.net/aqck2/4670.html

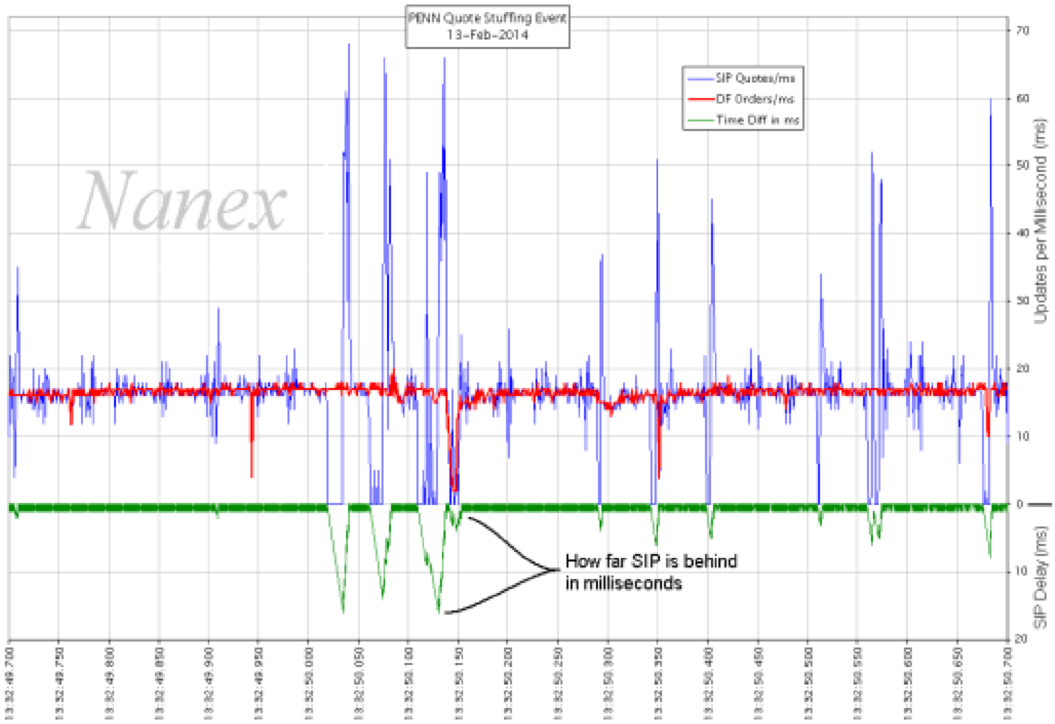


Figure 17: CDRG Quote stuffing 1

Source: Nanex

the fact that the arbitrage opportunities are created by the HFT strategy makes these opportunities more valuable than a traditional latency arbitrage that needs to be discovered by the HFT.

The chart reported in figure 18 analyzes the quoting activity of SIP and TotalView on a time interval divided in timeframes of 200 milliseconds. Over the considered time interval there are some timeframes where the SIP is silent, which means that it does not display any quotes; this is because of the Quote Stuffing strategy which congests the SIP that needs more time to process the overwhelming volume of orders with respect to TotalView; in particular, at 13:32:50:00 SIP displays the updated quotes with a delay of 16 microseconds with respect to TotalView[137].

Figure 19 shows the quoting activity on the Nasdaq from 2009 to the event of CDRG Quote Stuffing of February 2014. Blue dots identify a quoting activity of more than 6000 quotes per

---

[137] Id.

second, while the orange ones correspond to a quoting activity of more than 25 000 quotes per second, and the size of the dots is higher the higher is the number of quotes given these thresholds[138]. From 2009, the increasing sophistication of automated trading and HFT has increased the speed of trading activity: this is demonstrated by the abruptly increase over time of events recognized as quote stuffing events, with 6000 or more quotes per second. The CDRG Quote Stuffing strategy in February 2014 is easily visible in the high concentration of Quote Stuffing events with more than 25000 quotes per second on the right of the graph.

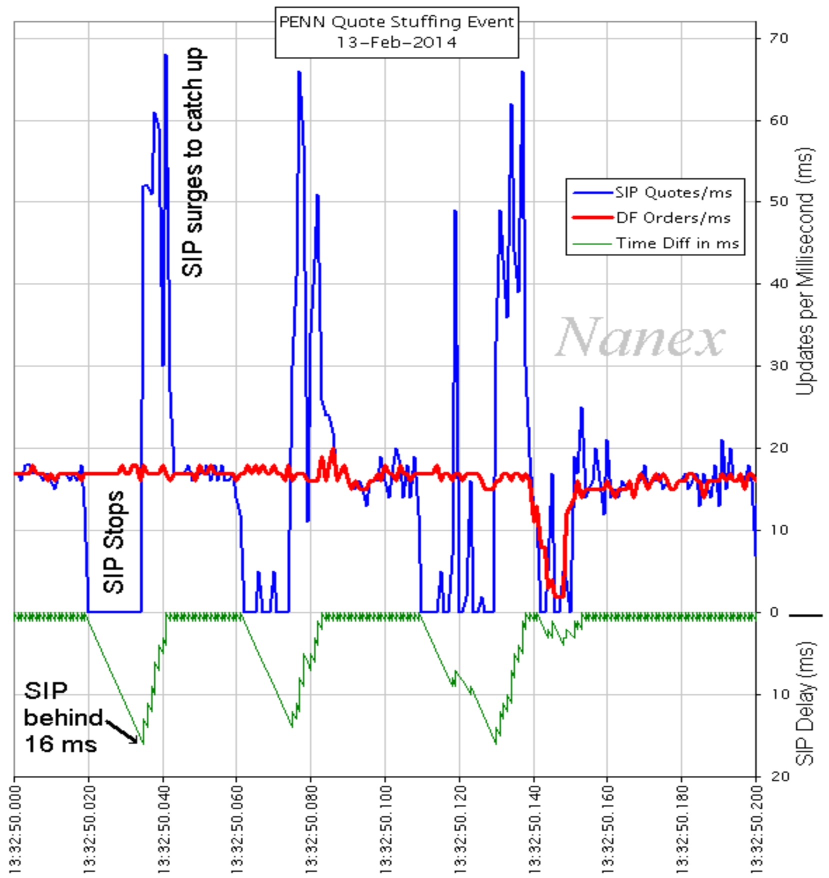


Figure 18: CDRG Quote stuffing 2

Source: Nanex

2.6 Trading on news

Given the high speed and frequency at which HFTs trade and their computational power, trading on news seems to be the more suitable strategy for this breed of traders. Trading on news means trading on the base of informations that are likely to affect the trend of financial instrument's prices: macroeconomics news such as unemployment rate or interest rate,

[138] Id.

companies earning announcements and so on; in particular, traders try to implement ever increasing sophisticated algorithms that are able to automatically use informations from a number of information providers as large as possible to implement pre-designed trading strategies. For instance, the use of algorithmic systems that are able to associate a particular trading strategy to a given pattern of words reported by different news providers: these algorithms allows the trader to significantly reduce the interpretation process of the news, reacting more quickly to material incoming news.[139]

Furthermore, the ultra-fast speed at which HFTs work allows them to fully exploit trading on news strategies: a successful news trader is able to react to new informations before other traders do; consequently, only the faster traders will be able to profit from this strategy, while the slower ones will trade on stale informations, already incorporated in the price.

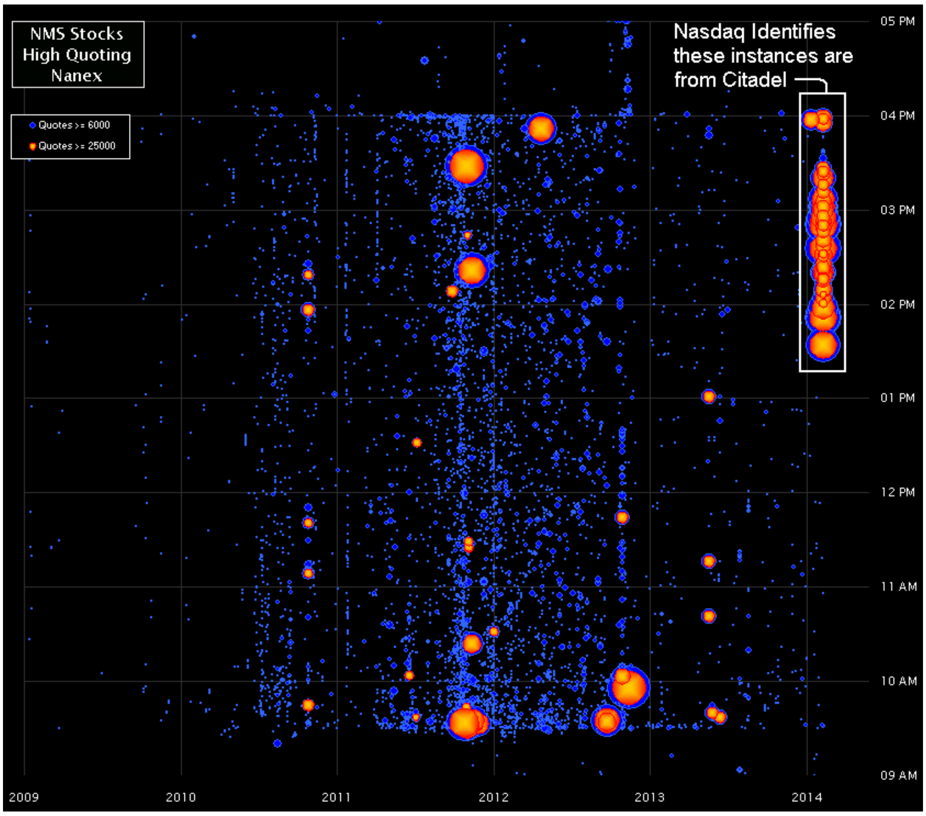


Figure 19: CDRG Quote stuffing 3

Source: Nanex

[139] Id. 120.

3 HIGH FREQUENCY TRADING AND MARKET QUALITY

The ever increasing market share of algorithmic trading and HFT has focused the attention of supervisory authorities and finance scholars on their effects on market quality, in terms of liquidity, volatility and efficiency, and in general on the systemic risk due to the contagion across different markets of a shock occurred in one of them and on the overall integrity of the market.

The main results of the academic literature on the impacts of HFT are not univocal: some studies have found beneficial effects of HFT on market liquidity, volatility and informational effciency, while others have pointed out how HFT has been detrimental for market quality, in particular during period of financial distress.

3.1 Liquidity

One of the most important indicator of market quality is the liquidity of the market: it can be defined as the ability to quickly trade large size at low cost; from this definition we can recognize three different dimensions of liquidity: immediacy, width and depth[140].

The immediacy refers to how quickly a trade can be executed: the higher the market is liquid, the more easier is for a trader to quickly find a counterparty to complete a round trip; in other words, when the market is liquid, a trader can quickly find another trader willing to trade with him.

The depth of the market, also known as market breadth, is the cost per unit of liquidity, the cost of doing a trade at a given size; for a retail trader the width is the bid-ask spread: given the small size of this kind of trade, it does not "walk the book" and it is immediately executed at the NBBO; thus, the cost of a round trip for a retail trader is equal to the quoted bid-ask spread.[141]

The depth of the market refers to the size of a trade that can be arranged at a given cost and it is meaured in units avilable at a given price of liquidity[142]. Thus, this liquidity dimension indicates the number and size of the resting limit orders pendant in the limit order book; an highy liquid market is characterized by a deep order book that is able to absorb large orders without generating significant market impact.

Finally, the last dimension of liquidity is the resilience of the market, that refers to how quickly the price of an asset reverts to its fundamental value after an order flow imbalance generated

[140] Id. 79.
[141] Id.
[142] Id.

by uninformed traders has caused the price to deviate from that value. If a market is liquid with respect to the first three dimensions, it will attract more traders, both uninformed and informed traders; if the latter are well capitalized and willing to trade, the deviation of prices from their fundamental value will be short-lived because informed traders will quickly correct the price inefficiency driving prices to their fundamental[143]. Figure 20 graphically represents the four dimansions of liquidity.

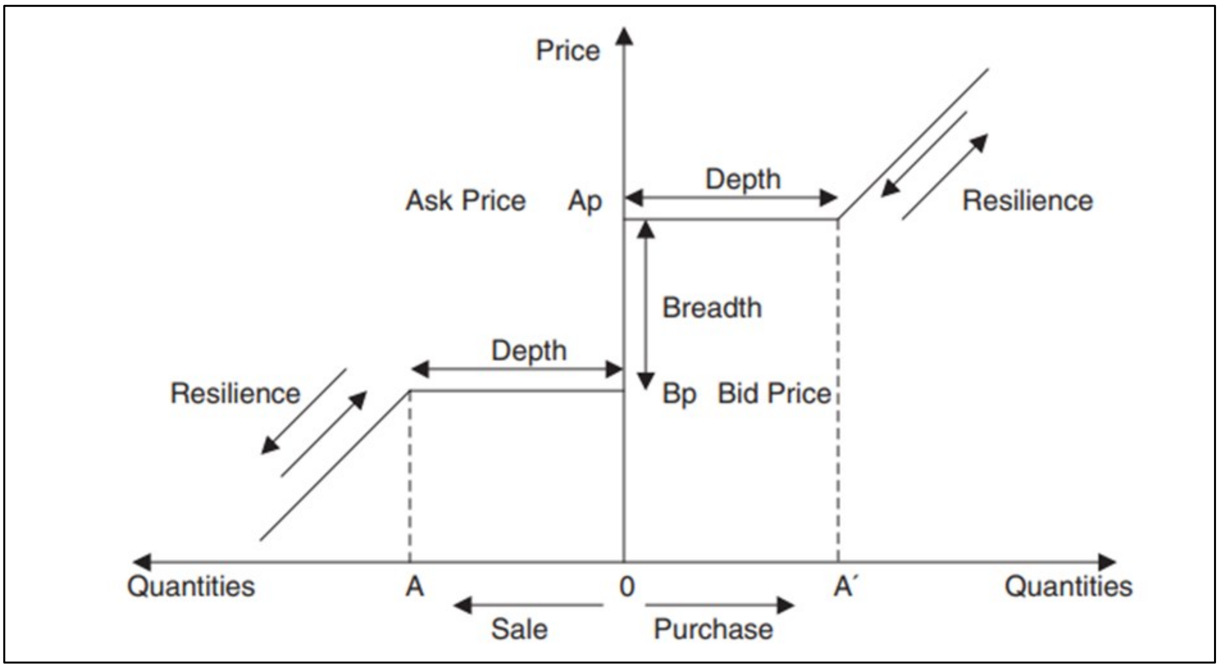


Figure 20: Liquidity dimensions
Source: Aldridge (2010)

3.1.1 HFT effects on market liquidity

A large part of academic literature on algorithmic trading and HFT have studied their effects on market liquidity and there is a strong evidence of a negative correlation between HFT activity and bid-ask spread.

As we have seen in section 1.3, Direct Electronic Access allows HFTs to trade on trading venues without the intermediation of a stock market's member, progressively replacing designated market makers, acting as informal market makers. The use of co-location services, direct market feed and automated algorithm enables HFT market makers to immediately react

143 Id.

to changing market conditions, continuously updating their quotes (with an higher order-to-trade ratio, characteristic of HFT). The speed at which HFTs are able to react to incoming market data significantly reduces the adverse selection risk to trade on the wrong side of the market with informed traders, that is the risk to suffer losses on the established position because of adverse price movements after the trade. Consequently, the higher is the adverse selection risk, the wider is the bid-ask spread set by market makers: they transfer the losses by trading with informed traders to uninformed traders, that will pay an higher transaction costs. From this reasoning, HFTs' technological advantage reduces the adverse selection risk and the quoted bid- ask spread.

On the other hand, the aforementioned causality can be on the opposite direction: HFTs are attracted by highly liquid market: given the high frequency at which they trade, a liquid market that allows HFTs to quickly move in and out of positions is a suitable market for this kind of traders. In order to avoid this endogenous problem (the problem that liquidity can be a cause and not a consequence of HFT activity), finance scholars studied the effect of HFT on market liquidity after the occurance of an exogenous shock that has affected the market quality through the increase of HFT activity.

Hendershott, Jones and Menkveld (2011) found a negative relationship between algorithmic trading and market liquidity; they studied the impact of algorithmic trading on different market liquidity measures after the introduction of Autoquote in NYSE in January 2003; the authors estimated algorithmic trading activity using the indirect method: the proxy for algorithmic trading is the number of electronic messages (order submissions, cancellations and executions) per \$100 of trading volume, useful to capture effects of an increase in the intensity of order submission and cancellations driven by algorithmic trading and HFT.[144] Figure 21 shows the evolution of the algorithmic trading proxy from 2001 to 2005: the trading volume for message traffic decreased for small, medium and drastically for large capitalization stocks; we can notice that the trading volume per message for large-cap stocks decreased from \$7000 per message in 2001 to \$1100 per message at the end of 2005.

They used the introduction of the Autoquote in NYSE as an instrumental variable[145] to solve the aforementioned reverse causality problem; as the name suggests, the Autoquote was an automated system that automatically updated inside quotes whenever the arrival of a new order changed the limit order book. Clearly, this was a market structure change that affected

[144] Hendershott T., Jones C. M., Menkveld A. J. (2011), Does Algorithmic Trading Improve Liquidity?, The Journal of Finance Vol. 66, No. 1, 1-33.

[145] In econometrics, an instrumental variable is a variable that satisfies the exclusion restrictions: it is correlated with the endogenous variable but not with the dependent variable; thus, it is a variable that affect the dependent variable only through the effect of the endogenous variable.

algorithmic trading activity, not the trading behavior of slower traders, because differenty from the slower-reacting humans, algorithmic traders have the technological advantages to quickly react to the more immediate feedback about market conditions provided by the Autoquote system.[146] The authors found that an increase in algorithmic trading activity reduced the quoted bid-ask spread; in particular, a decrease of $100 of volume per message narrows the quoted bid ask spread by 0.53 basis points.[147] The result is statistically significant only for large capitalized stocks, not for medium and small size stocks; this is consistent with the use of algorithmic trading mainly for large size stocks: thus, the introduction of NYSE Autoquote may have had little or no effect on less liquid stocks.

They also found that the same increase in the algorithmic trading proxy led to a reduction of market depth by about 5% with respect to the daily average of the period from December 2002 to July 2003; this result is consistent with the small size trade usually executed by HFTs and with the practice of using algorithmic traders to "slice and dice" large orders to minimize the market impact[148].

The authors also anayzed the effect of algorithmic trading on other liquidity measures: the effective spread and the realized spread. Indeed, the quoted bid-ask spread represents the transaction cost only for small trade, whose size does not exhaust the available liquidity at the NBBO; the quoted bid-ask spread is defined as follow:

$$QS_{it} = 100 * (Ask_{it} - Bid_{it})/M_{it}$$

Where Ask_{it} and Bid_{it} are respectively the ask and bid price of security i at time t and M_{it} is the quote midpoint between the ask and bid price, that is usually used to identify the fundamental value of security i at time t[149].

However, the effective spread cannot represent the transaction cost for large orders because they are not entirely executed at the bid-ask quote, but they "walk down the book" when the broker improves the execution price at inside quotes, even undisplayed, or "walk up the book" when they are executed at inferior prices when the liquidity exhausted progressively. Thus, the effective half spread is equal to:

$$ES_{it} = 100 * D_{it} * (P_{it} - V_{it})/V_{it}$$

[146] Id. 144.
[147] Id.
[148] Id.
[149] Hendrick Bessembinder, Kumar Venkataraman, 2009, Bid-Ask Spreads: Measuring Trade Execution Costs in Financial Markets, Encyclopedia of Quantitative Finance.

Where P_{it} is the transaction price for security i at time t, D_{it} is an indicator variable that is equal to 1 for buy orders and -1 for sell orders and V_{it} is the fundamental value of the security or the benchmark price that an algorithm has to meet[150].

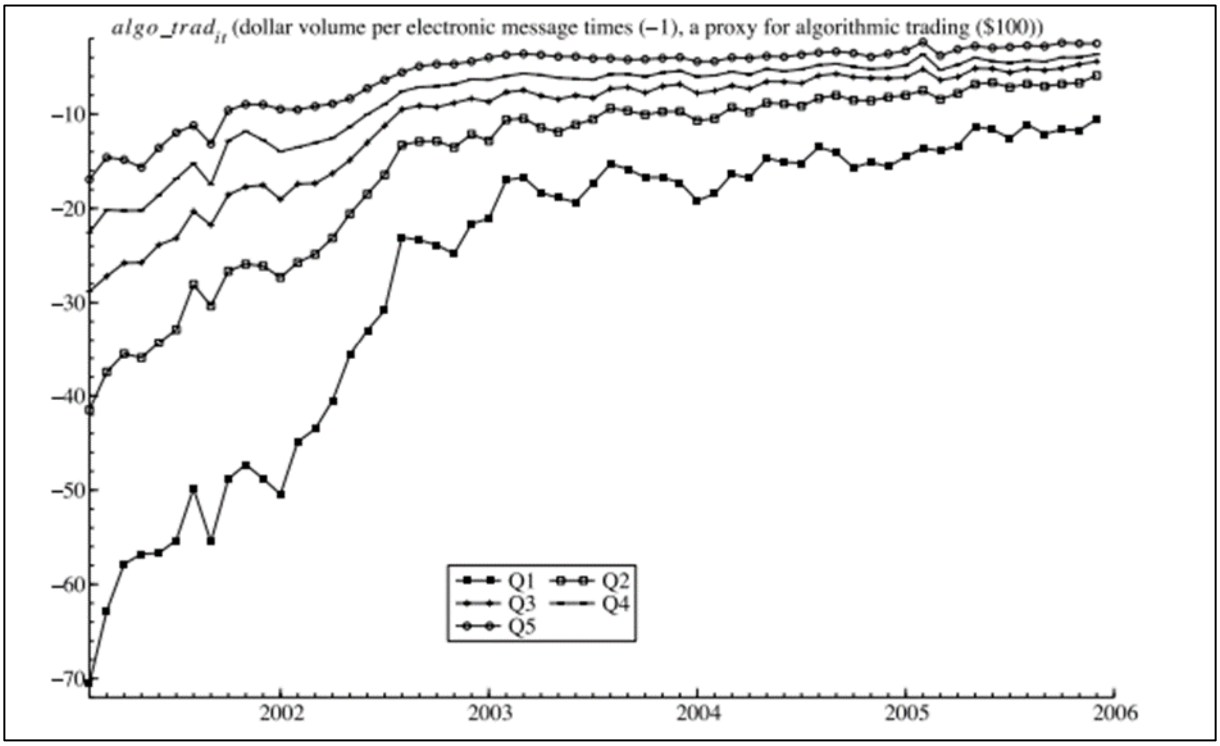


Figure 21: Evolution of algo trading proxy

Source: Hendershott, Jones, Menkvekd (2011)

The effective spread is defined as the gross revenue of market makers; the net revenues of market makers must take into account the price impact of the order when it trades with informed traders that is equal to:

$$Price\ impact = 100 * D_{it} * (V_{it+n} - V_{it})/V_{it}$$

Where V_{it+n} is the price adjustment of the stock traded after the private information of the informed trader has been incorporated in the price[151]. The price impact indicates the reduction of the effective spread when the market maker trades with an informed traders: it buys when

[150] Id.
[151] Id.

the asset is overvalued and sell it when it is undervalued. Thus, the net revenues of market makers, the realized spread, is the difference between the effective spread and the price impact. The realized spread is also defined as follows:

$$RS_{it} = 100 * D_{it} * (P_{it} - V_{it+n})/V_{it}$$

Hendershott et al. (2011) found that algorithmic trading has also reduced the effective spread for medium and large-cap stocks; then, when they decomposed it in realized spread and price impact, they found that algorithmic trading activity has increased realized spread, but this was offset by a reduction in market impact; this means that the reduction in the effective spread paid by traders is not due to the capacity of algorithmic traders to compete against designated market makers, reducing aggregated market makers' revenue, but instead on the capacity to reduce the adverse selection losses from informed liquidity demanders[152].

Similar results have found by Brogaard (2011); the author artificially created two order books, one for the quotes posted by HFTs and the other for the quotes posted by non-HFT for a sample of 120 stocks, 60 NYSE and 60 Nasdaq listed stocks, each group equally divided in small, medium and large capitalization stocks, over two years of trading from 2008 to 2009. The data on quotes are from Nadaq and BATS and the HFT identification method used is the direct one: HFT firms are those flagged by Nadaq and BATS as firms whose core activity is high-frequency trading. For Nasdaq data he found that HFTs matched or improved the non-HFT inside quotes 65% of the time, while for BATS HFTs matched or improved them 55% of the time. It also demonstrated how HFTs are more active on large-cap stocks: on Nasdaq, they improve quotes 83% of the time, against an improvement rate of 51% for small cap stocks.[153] Menkveld (2013) found a direct evidence of a casual link between the activity of an HFT market maker and Dutch index stocks after the launch of Chi-X, a pan-European trading platform operated by Instinent. HFT market makers were indirectly identified by the author: it used resting limit orders in 80% of its trades on dutch stocks, its net position on them crossed zero several times during the trading day both in Chi-X and in the incumbent stock market Euronext, and it usually had flat end-of-day-positions. Chi-X began to trade dutch stocks on April 2007, but the market share on those stocks significantly rose to double digit only after the entry of the HFT market maker on August 2007. Shortly after its entry, the HFT market maker became the major trader on dutch stocks on Chi-X, with a participation rate between 70% and 80% of trades. Since Chi-X began trading dutch stocks one year prior to belgian stocks, the author

---

[152] Id. 144.

[153] Brogaard J., 2011. High frequency trading and market quality. Working Paper, University of Washington.

compared the quoted bid-ask spread of dutch stocks with that of belgian stocks, traded on Euronext, after the entry of the HFT market maker to study the effect of it on market liquidity. After the entry of the HFT market maker, the treated group (the Dutch stocks) bid-ask spread decreased by 50% with respect to the one of the untreated group (the Belgian stocks): this decrease, in coincidence with the high participation rate of HFT market makers in Dutch stocks on Chi-X, is interpreted as a negative relation between HFT market making and bid-ask spread[154].

3.2 Phantom liquidity

Despite the aforementioned strong evidence of the beneficial effects of HFT market makers, criticism of HFT refers to the liquidity provided by HFTs as phantom liquidity; as we have seen in section 2.5 , HFTs engaging in ghost liquidity strategies submit “flickering quotes”, resting limit orders that last on the limit order book for a few milliseconds before being cancelled, without the intention to execute them, but with the aim to give other traders misrepresentation of the market liquidity. Thus, HFTs have significantly reduced the informational power of the order book with a liquidity that is only apparent, making more difficult for investors to evaluate whether or not the posted liquidity is transient and thus to formulate an optimal execution strategy[155]. As a result of this illusory liquidity, the transaction costs of traders can increase significantly; to clarify how HFTs’ phantom liquidity can hurt traders let assume a trader whose trading strategy is based on market microtrends, which means that it trades on the base, for instance, of the trading behavior of other traders, focusing its attention on how the order book changes. An HFT that is aware of this trading behavior can simulate particular liquidity dynamics to instigate trader’s action; the HFT can “ping” the market with a series of limit sell orders at decreasing limit prices to give the trader the impression of a selling pressure, gradually empting the bid side depth of the order book.[156] As a consequence, the trader initiates to sumbit sell orders; the HFT fills these orders establishing a long position and shortly after it cancels the other limit sell orders and initiates the reverse strategy, “pinging” the market with buy orders to artificially create a buy pressure. The trader is forced to liquidate its position to not incur large losses given the price increase, in particular if other traders initiate to buy, and the HFT

[154] Menkveld A., 2013, High frequency trading and the new-market makers, Journal of Financial markets, Vol 16 (4), 712-740.
[155] Id. 122.
[156] Decreasing number of limit buy orders in the order book is associated with a decrease in the price of the financial instrument because market sell orders associated with a price drop have the effect to decrease the bid-side depth because they fill resting limit buy orders or because traders tend to cancel their limit buy orders during a selling pressure. Furthermore, a tighter buy side depth increases the probability of further price decrease because it is not able to absorb large sell orders without significantly decreases the price of the stock.

will fill the orders of the trader at a gain[157]. Thus, the phantom liquidity generated by the HFT has forced the trader to trade when it would have not traded, closing its position at a loss.

3.3 Volatility

Volatility is a fundamentally important concept in finance: it is a measure of the variability of a security's return over time. Conventionally, the volatility of stock X, called historical volatility, is determined by the standard deviation of stock's return, σ_x, over a period T:

$$\sigma_x = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(R_i - \bar{R})^2}$$

Where the standard deviation is the square root of the variance of stock X, which measures the variability of stock X's return over period T as the mean of the square of the deviation of stock' return R_i from the mean return over T, $\bar{R}$. Thus, volatility is a measure of the dispersion of an asset's return from its expected value.

Volatility is strictly associated with risk: the higher is the volatility of an asset's return, the higher is the probability the asset will assume very high or low values; in the first case a significant increase in price generates losses for traders who have short positions in the asset, while significant drops in price will hurt traders with long positions.

Volatility is also strictly correlated with market liquidity: the higher is the volatility of an asset, the wider will be the bid-ask spread charged by market makers; an highly volatile asset is more likely to largely deviate from its mean value, a value that is usually used by market makers as the midquote of the posted bid-ask spread. Thus, the higher is the asset's volatility, the higher is the probability that the asset's price will move against the inventory position of the market maker: it rapidly and significantly increases after the market maker has sold the asset, or it significantly and rapidly drops after he has bought the asset. Consequently, the market makers charge wider bid-ask spread on volatile assets in order to reduce the inventory risk.

[157] Id. 97.

3.3.1 HFT Effects on market volatility

Academic literature's findings on the effects of HFT on market volatility is not unambiguous; as we have seen in the previous section, when HFTs act as market makers they reduce market volatility: by providing liquidity, they reduce the probability of large price jumps or drops along the limit order book. Hagströmer and Nordén (2013) studied the relationship between HFT market making activity and bid-ask spread on the Nasdaq-OMX Stockholm exchange (NOMX-St). From the relation between HFT activity and market liquidity seen in the previous section, we can deduct that also for volatility the direction of causality may be reversed: since HFTs are more active in liquid stocks, low volatile assets attract HFT activity. After having controlled for the endogenous problem studying the relationship between the two variables after tick size changes, the authors found a negative correlation between HFT market making activity and the short-term volatility of the 30 most traded stocks on the NOMX-St, those that constitute the leading Swedish stock index OMX 30[158]. Hasbrouck and Saar (2013) found similar results for Nasdaq listed stocks both for a relatively flat market period, September 2007, and a period of higher uncertainty in the market, June 2008, between the first sale of Bear Stearns in March and the Chapter 11 filing of Lehman Brothers in September; for both periods, the authors found that an increase in low-latency trading has decreased short term volatility. [159]

The HFT effect on market volatility changes if we consider aggressive HFT strategies instead of passive strategies. HFTs engaging in momentum ignition strategies increase price volatility: instigating other traders to trade in the HFTs' desired direction causes prices to fluctuate more than would have done if HFTs had not implemented such a trading strategy. Electronic front running and low-latency arbitrage strategies increase price volatility: once HFTs has become aware of an incoming large orders after having succesfully "ping" the market or throught the use of direct market feed or co-located servers, they trade ahead of the large order, reducing the available liquidity and consequently increasing price movements when the large order will be executed.

Boehmer et al. (2012) found an international evidence of the positive relarionship between HFT activity and market volatility; they studied HFT effect on volatility for 39 stock markets around the world, excluding US, from 2001 to 2009 for a sample of 12.800 stocks. They used as

[158] Hagstromer, B. and Norden, L., 2013, 'The diversity of high frequency traders', Journal of Financial Markets, 16(4), 741-770. The authors use as exogenous shock for HFT market making activity tick size changes; in many European stock exchanges the tick size on a stock changes when the price of it crosses a particular threshold; for instance, in the NOMX-St, the tick size for stocks priced below SEK 50 is SEK 0,01, while for stocks price between SEK 50 and SEK 100 is SEK 0,05. Since market maker's profit increases with wider bid-ask spread, increase in tick size increases HFT market making activity.

[159] Hasbrouck, J. and Saar, G., 2013, 'Low-Latency Trading', Journal of Financial Markets, Vol. 16 (4), 646-679.

exogenous shock for HFT activity co-location services introduction for each stock market; even if they found that an increase in HFT activity increased market liquidity and price informational efficiency, they also found a positive correlation with short-term market volatility[160]. Futhermore, in this paper the authors distinguished two types of volatility: a good volatility and an excessive volatility; the good volatility is the desiderable component of volatility: this volatility is generated whenever HFTs quickly trade on new informations. Thus, desiderable volatility is associated with faster price discovery; on the other hand, the excessive volatility, also known as noise or transitory volatility, does not contribute to market efficiency. Excessive volatility is generated by uninformed traders, whose trading move prices away from their fundamental value, or by HFTs, when their trading is not driven by fundamental information but on the strategy to quickly liquidate their position.[161] With respect to their study, Boehmer et al. (2012) found that HFT activity increases excessive volatility.

Valeria Caivano (2015) studied the effect of HFT on the italian market volatity; she studied the impact of 14 HFT firms operating in the italian market according to the ESMA after the migration of Borsa Italiana cash market to Millenium Exchange, an ultra-fast electronic trading platform that significantly reduced market participants' trading latency. The author found a positive impact of HFT activity on the volatility of 39 italian stocks; in particular, they estimated that an increase by 10% of HFT activity increases intraday voatility between 4% and 6%. The result did not significantly change, with an effect between 3% and 5%, when the author considered a larger dimension for HFT activity, which takes into account not only the activity of the 14 HFT firms, but also HFT activity carried out by firms whose core activity is not HFT, such as investment banks[162].

Concerns on HFT effect on volatility derived not only from the aforementioned aggressive strategies, but also from the lack of dependability of liquidity provided by HFT market makers[163]. Although HFT market makers reduce volatility during relatively flat markets providing cheaper liquidity, this beneficial effect disapperars during period of high volatility. High volatility increases market maker's inventory risk; since HFT market makers are not subject to the exchange-registered market makers' affirmative obligation to continuously quote two-sided markets of the securities they quote, HFTs can quickly withdraw their liquidity, with the consequence to exacerbate market volatility.

[160] Ekkehart Boehmer, Kingsley Fong, Julie Wu, 2012, International evidence on algorithmic trading, Working Paper, EDHEC Business School.
[161] Id.
[162] Valeria Caivano, 2015, The impact of High-Frequency Trading on Volatility: Evidence from the Italian Market, Quaderni di Finanza n° 80, Consob.
[163] Id. 122.

One of the most important event that pointed out the HFTs' role in contributing to exacerbate extreme market movements was the May 6, 2010 Flash Crash.

3.4 May 6, 2010 Flash Crash

Despite some empirical researches have demonstrated the negative correlation between HFT activity and market volatility, the same result is not found in periods of stressed market conditions, when HFT has exacerbated financial instability, contributing to generate unfairly and disorderly markets. The May 6 2010 Flash Crash was the first and most important event that shed light on the destabilizing behaviour of HFTs in distressed markets: on that date stock market indexes experienced a severe price decline followed by a very quick recovery; the Down Jones Industrial Average (DJIA), a stock index market participants refer to as benchmark for the overall market performance, fell by 998.5 points, almost the 9% of its value, experiencing the largest intraday decline of its history; differently from previous market crashes, after the DJIA has reached the daily lowest value, it rebounded to the pre-crash level in a few minutes. Similarly, the E-mini S&P500 futures and the S&P500 Stock index dropped by 5% in less than six minutes before rising to the pre-crash prices in just over twenty minutes (Figure 22)[164].

Figure 1: Select Equity Indices and Equity Index Futures, May 6, 2010

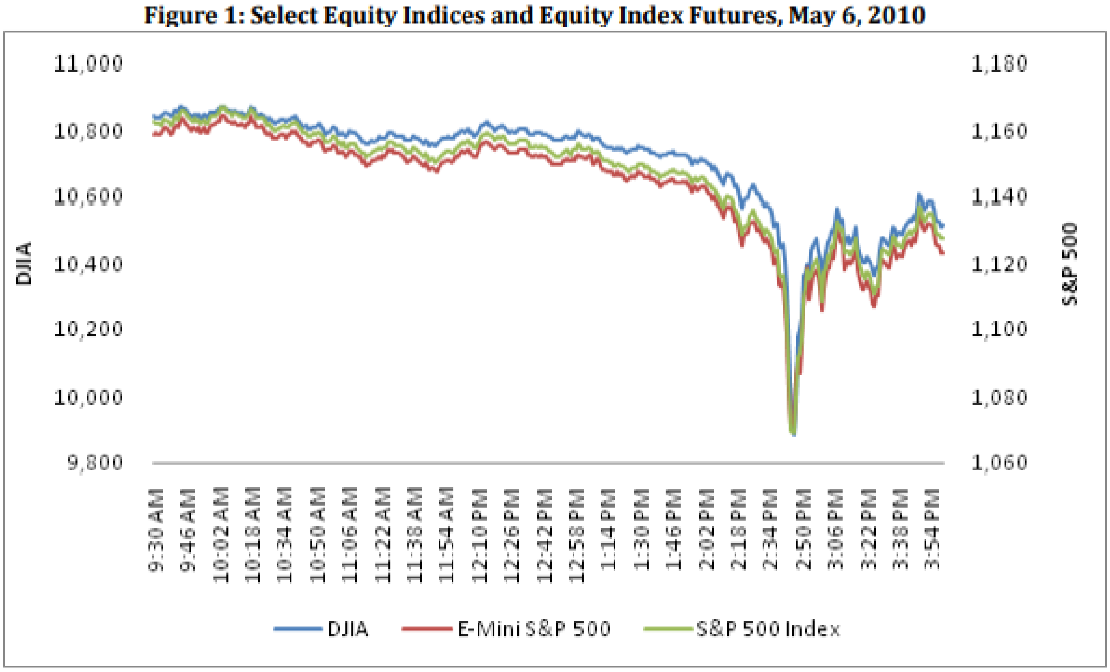


Figure 22: Select Equity indexes and Equity Index Futures, May 6, 2010
Source: Preliminary Findings Regarding the Market Events of May 6, 2010.

---

[164] Report of the Staffs of The CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues (September 30, 2010), Findings Regarding the Market Events of May 6, 2010.

3.4.1 Pre-crash market condistions

May 6, 2010 started as a day of high uncertainty for financial markets; concerns for Greek debt crisis generated speculation and fear that it would ignite a string of defaults across Europe and the rest of the world. The uncertainty of that day was reflected by several market data. Before the Flash Crash, concerns from oversea generated an early selling pressure that drove major equity indexes and equity index futures into negative territory: for instance, the DJIA declined 161 points (-1.5%) from the opening of the market to 2:00 p.m; the S&P 500 Index declined 33 points (-2.9%) and the E-mini S&P 500 Index June futures lost 15 points (-1.3%) in the same period[165].

Conventionally, the most reliable indicator of market uncertainty investors refer to is the Chicago Board Option Exchange SPX Volatility Index ("VIX"); also known as "fear index" or "investors fear gauge", the VIX estimates the expected volatility of the S&P 500, the U.S. most importan stock index, for the next thirty days, from the implied volatility[166] of the call and put options on the S&P 500 index; the VIX is a weighted average of the implied volatility derived from the price of the call and put options on the S&P 500 index; since this index uses the implied volatility of the options on the largest stock index, it can be interpreted as a proxy for market participants' expectation on the future market volatility.

Although the VIX only measures the magnitude of the expected market volatility, not its direction, from the historical negative correlation between the VIX and the S&P 500 index we can conclude that an high VIX reflects negative expected market returns.[167] On May 6 2010 the opening VIX was at 25,88, 15,5% higher than its value of 22,41 at the beginning of the week; the VIX steadily increased during the day, reaching its highest value by 2:46 p.m. at 40,26, with an increase from the morning level equal to 55,56%, signalling a significant increase in market uncertainty. When the market closed at 4:00 p.m. the VIX was at 32,80, 31,7% higher than the previous closing day's value[168].

The market uncertainty of that day was reflected by a "flight to liquidity": the bearish expectations on market returns drove investors' liquidity from riskier assets, such as stocks, to safer ones, such as bonds or gold. On May 6, 2010 the price of gold futures contract rose from $1180 to $1210 for troy ounce; Furthermore, the ten-year U.S. Treasury yield declined from

---

[165] Report of the Staffs of The CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues (May 18, 2010), Preliminary Findings Regarding the Market Events of May 6, 2010.

[166] The implied volatility is derived from option pricing model, such as the Black-Scholes model, and it is a metric used by investors to estimate the option's underlying stock's future price movement; in other words, it is the expected volatility of the option's underlying asset.

[167] Dondoni, A., Maggi, M., & Montagna, D., 2018 "VIX Index Strategies Shorting volatility as a portfolio enhancing strategy". Universita' degli Studi di Pavia, Banca IMI.

[168] Id. 165.

the previous closing value at 3,58% to an intraday lowest value of 3,26%[169].

The spread on Credit Default Swap on Greek government bonds rises to 937,9 basis points from the previous closing day value of 844,2 basis points: an increase in the CDS spread means that the buyer of the CDS contract, the one who want to be insured against the default of the Greek debt, has to pay an higher premium to the seller of the CDS contract, the one that provides the insurance against the payment of the premium, because of the increase of the default probability of the Greek goverment bonds[170].

On the wave of the Greek debt crisis, concerns on the deterioration of European Union financial stability generated a downward pressure of the Euro on the currency markets, with a significant decrease in the EURO/USD and EURO/JP YEN exchange rate[171].

Finally, the turbulence on that day was also signalled by the triggering of a large number of NYSE Liquidity Replenishment points (LRPs). LRPs are control systems adopted by the NYSE to dampen volatility on securities whose price movements have exceeded a pre-specified price band, usually from 1% to 5% of share price. Once the LRP on a certain stock is triggered, the market on that stock temporarily "go slow", shifting from being automatic to manual, allowing market makers to provide additional liquidity to stabilize prices within the LRPs' limits before market to return automatic. On May 6, 2010, the number of LRPs triggered on the NYSE for securities listed and traded on the exchange was higher than the average levels, which reflected an higher volatility in the market from the morning.[172]

3.4.2 The execution of the "Sell Algorithm"

The SEC-CFTC joint report on the May 6, 2010 Flash Crash recognized as the main cause of the crash the execution of a large sell order in the E-mini S&P 500 futures market by Waddell & Reed Financial, Inc, a complex mutual fund, over the Chicago Mecantile Exchange (CME) in a very short period. At 2:32 p.m the mutual fund begans the execution of a 75000 E-mini futures contracts sell order (approximately $4.1 billions) with the aim to hedge an existing long position on the underlyng S&P 500 stock index through the deployment of a computer algorithm that was programmed to feed the large order in the June E-mini futures market with a patecipation rate equal to 9% of the trading volume in that market in the previous minute[173]. Differently from the algorithms that we have seen in section 1.5, with VWAP algorithm or

[169] Id.
[170] Id.
[171] Id.
[172] Id.164.
[173] Id.

Implementation Shortfall algorithm that have as inputs of their strategies volume, time and price in order to minimize the overall investors' transaction costs with respect to a pre-determined price benchmark, the algorithm used by the mutual fund did not take into account any of these variables. Thus, the execution of an algorithm designed in these terms would have not prevent the large sell order to be executed at extremely unfavourable conditions: without targeting the time over which execute the order or the average execution price to reach, the large sell order would have been subject to large market impact due to fast execution or to unfavourable prices in period of lack of liquidity. Months before May 6, 2010, the same trader executed an order of the same size using a combination of automated and manual trading strategy that on the contrary have taken into account volume, time and price; in that occasion, the large order was executed in 5 hours, while on May 6, 2010 it was executed in 20 mnutes with an expected abrupt price decline of the june E-mini futures, which declined by 5% in five minutes[174].

Before studying the role of the HFTs in the Flash Crash we examine the market participants that interacted with the large sell order.

3.4.3 Market participants

Following a study of Kirilenko, Kyle, Samadi, and Tuzun (2017), market participants are classified into 6 categories: intermediaries, HFTs, Fundamentl buyer, fundamental seller, small traders and opportunistic traders[175].

Among the 15000 accounts that traded in the E-mini futures market from may 3 to May 6, 2010, intermediares are those traders that provide liquidity quoting two sided market; thus, they consistently buy and sell during the trading day while maintaining a low intraday level of inventory.

According to the trading pattern of an intermediary, a trader must satisfy three criteria with respect to the daily trading volume, the intraday inventory pattern and the end-of-day net position.

- With respect to the trading volume, the authors recognize as intermediaries in the E-mini market the accounts that have traded at least ten E-mini contracts in one or more of the three days prior to May 6:

[174] Id.

[175] Andrei Kirilenko, Mehrdad Samadi, Albert S. Kyle, Tugkan Tuzun, 2017, The Flash Crash: The Impact of High Frequency Trading on an Electronic Market, The Journal of Finance, Vol. 72, No 3, 967-998.

$$Vol_d \geq 10,$$

where Vol_d is the number of E-mini contracts traded by account d in one of the three day prior to May 6. Since intermediaries provide liquidity to traders who demand immediacy, they have to participate in an high number of transactions. This cut off level allows authors to not erroneously classify as intermediaries small traders, that are traders who trade an insignificant amount of contracts; however, in order to not classify as intermediaries large buyers or sellers, they added two other criteria that identify the trading pattern of the intermediary[176].

- With respect to the intraday inventory pattern, the three-day average of the absolute value of the ratio of the account's end-of-day net position to its daily trading volume must not exceed 5%:

$$\frac{\sum_{d=1}^{3} \frac{|NP_{d,t=405}|}{Vol_d}}{3} \leq 5\%$$

Where, once the trading day was divided into 405 minutes, $NP_{d,t=405}$ is the net position (long position minus short position) of the account d at the end of the trading day. The market making role of intermediaries imply that they do not carry out significant overnight position[177].

- Finally, with respect to the intraday inventory pattern, the three day average of the square root of the account's daily mean of squared end-of-minute net position deviations from its end-of-day net position over its daily trading volume must not exceed 0.5%

$$\frac{\sum_{d=1}^{3} \sqrt{\frac{1}{405} \sum_{t=1}^{405} \left(\frac{NP_{d,t} - NP_{d,t=405}}{Vol_d}\right)^2}}{3} \leq 0.5\%$$

Where $NP_{d,t}$ is the minute by minute net inventory position of account d. It is a measure of the mean deviation of the minute by minute account's inventory position with respect to the end-of-the position during the three days prior to May 6[178]. This last criteria identifies the mean

[176] Id.
[177] Id.
[178] Id.

reversion behavior of the intermediary's inventory position: clearly, the market making strategy does not allow the net holding to accumulate a large position in each direction. HFTs are identified as intermediaries that trade more frequently during the sample period; thus, the 16 intermediaries that traded the largest number of contracts are classified as HFTs, the others as market makers. From may 3 to may 5 the 16 HFTs participated in an amount of trades 30 time higher than 189 market makers: the HFTS' average daily number of trades per second is equal to 5.89, against the 2.14 of 189 marekt makers.

Small traders are those that trade less than ten contracts each day, while large buyers and sellers are identified as traders that trade more than 10 E-mini futures contracts each day and whose net long end-of-day position and net short end-of-day position is at least the 15% of their daily trading volume. Finally, opportunistic traders are traders that do not fall in any of the above categories; in particular, they are traders that follow trading strategies such as cross market arbitrage, statisical arbitrage and news arbitrage[179].

3.4.4 Liquidity crisis in the S&P 500 E-mini futures market

At 2:32 p.m. the "Sell Algorithm" initiated to feed limit buy orders in the June E-mini S&P futures market; this contract is a small size option of the S&P 500 futurs, introduced by CME on the CME Globex trading platform in 1997. The market value of an E-mini S&P 500 futures contract is equal to one fifth of a standard S&P 500 futures contract and it is equal to $50 times the value of the S&P 500 stock index (the S&P 500 futures contract's notional value is $250 times the level of S&P 500 index); for instance, on May 6, 2010, the S&P 500 index was at 1100 points, thus the notional value of the E-mini futures was equal to $55000. The tick size for this contract is equal to 0,25 index points or $12,50 (since for the E-mini contract one index point is equal to $50, then the minimu tick size is equal to $0{,}25x\$50 = \$12{,}50$)[180].

On May 6, 2010, on the wave of an increasing market uncertainty, the buy side liquidity of the E-mini futures started to deteriorate since the morning; figure 23 displays the buy side and sell side market depth of the E-mini futures during May 6, computed as the sum of all the resting limit orders on both sides of the E-mini futures market on the CME; the buy side and sell side market depth began to diverge from 10:00 a.m. and by 2:00 p.m., almost 40 minutes before the Flash Crash, the sell side market depth was already twice the buy side one (160000 resting contracts on the sell side of the electronic order book against 80000 resting contracts on the buy side) and the buy side market depth fell to 73% of the morning average buy side liquidity (from

[179] Id.
[180] Id. 164.

a morning average between 9:30 a.m. and 10:00 a.m. of $6 billions to $4,4 billions) ; this pattern on the E-mini futures market highlighted an increasing selling pressure during the day as an hedging strategy against bearish market expectations[181].

From 2:32 p.m. to 2:41 p.m. HFTs and intermediaries were the first buyers of the first portion of the Sell algorithm's order; HFTs established a net long position equal to 3300 E-mini contracts. However, given the HFT strategy of not accumulating large intra-day net holding position, they started to liquidate the long positions on the E-mini, aggressively selling between 2:41 p.m and 2:44 p.m. 2000 E-mini contracts in order to re-balance their inventories[182]. Futhermore, from 2:32 p.m to 2:45 p.m fundamental sellers sold 80000 E-mini contracts, a traded volume 15 times larger than the one traded on average over the same 13 minutes interval in the

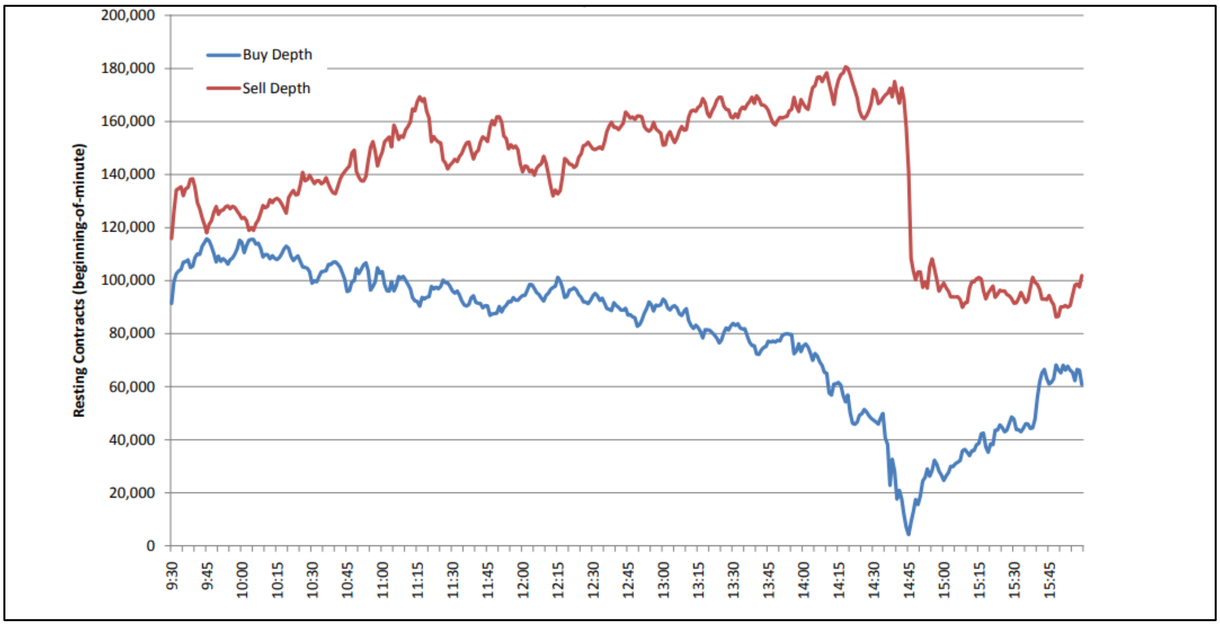


Figure 23: E-Mini Buy-Side and Sell-Side Market Depth

Source: Findings Regarding the Market Events of May 6, 2010

previous three days; the selling pressure generated by HFTs and fundamental sellers brought E-mini price down by 3% from 2:41 p.m to 2:45 p.m., with a decline from 1113 to 1056 points. During the same period fundamental buyers and opportunistic traders looking for cross-markets

[181] Id.

[182] See note 127: during the three days prior May 6, 2010, HFTs did not accumulate inventories greater than 4000 E-mini contracts in either direction.

and cross-products arbitrage opportunities bought E-mini contracts, but not in sufficient quantity nor at a fast enough pace to keep up with the selling pressure in the E-Mini, consequently generating a liquidity mismatch between buy side and sell side[183].

The period between 2:41 p.m. and 2:45 p.m. was also characterized by an higher level of "hot potato effect" compared to the previous three days; the rapid and significant decline of E-mini price breached market maker's and fundamental buyer's control system's internal risk limits that triggered price-driven integrity pauses, automated system that pause trading when market data appear to be questionable. Consequently, the buy side withdrawal of liquidity resulted in an HFT-to-HFT trading, that quickly buy and sell one another thousands of E-mini contracts: between 2:45:13 to 2:45:27, within only fifteen seconds, HFTs traded over 27000 E-mini contracts, while buying only about 200 additional contracts net, causing the E-mini price to further drop by 1,7%[184].

At 2:45:28 p.m. the E-mini buy side liquidity dropped to 1050 contracts, or $58 millions, less than the one percent of the morning average buyside market depth; The E-mini futures reached the intraday low ot 1065,79 (Figure 24) and the bid-ask spread in the E-mini S&P 500 market widened 6,5 points (26 ticks)[185]. The large widening of the bid-ask spread in the E-mini futures triggered for the first time in 2010 the CME "Stop logic" functionality, a circuit breaker designed to prevent the execution of stop loss orders that would have exacerbated the cascade in prices outside a "no bust" range, with the domino effect resulted by one stop orders triggering others; in the case of E-mini futures, the "no bust" range is equal to 6 index points (24 ticks) in either direction. On May 6, 2010 CME "stop logic" functonality curbed the abrupt downward generated by the Flash Crash halting trading in the E-mini futures market for an interval of 5 seconds, the so called stop logic reserve period, from 2:45:28 p.m. to 2:45:33 p.m. The reserve period allowed market participants to better evaluate the trading that has occurred, giving them a more transaparent and organized opportunity to offset the order imbalance that has caused the volatility; during this period traders can submitted, modified or cancelled orders, but no executions could take place; once the reserve period was finished and traders have rationally reassessed their valuation on the E-mini futures market price, trading reopened with a call auction to determine the starting price before returning to the standard continuous trading [186].

[183] Id.
[184] Id.
[185] Id.
[186] In a call auction market, buy and sell orders are collected together over a period during which traders express their trading interests and executed at the same time at a single clearing price. The single clearing price is equivalent to the best bid and offer if they coincide, or their mean if they are different; all the bids above the clearing price and all the asks below it are executed at the clearing price.

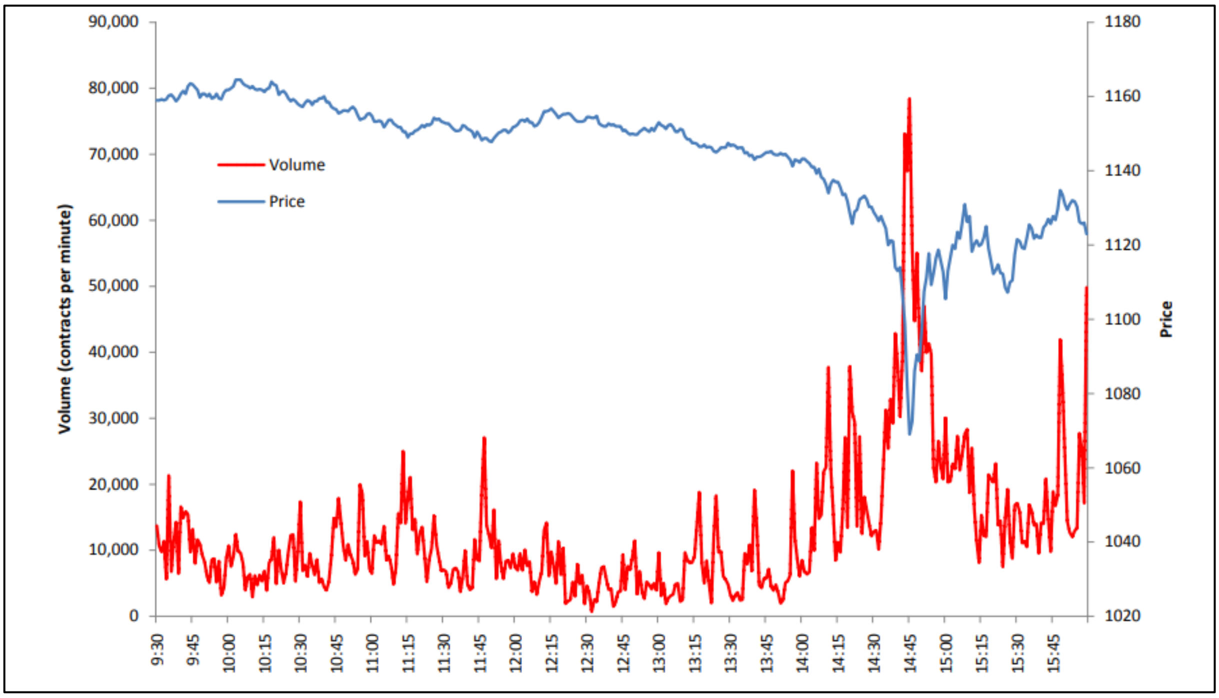


Figure 24: E-Mini Volume and Price

Source: Findings Regarding the Market Events of May 6, 2010

Figure 24 also shows that the rapid decline of the E-mini price was accompained by a large increase in the trading volume; in particular, from 2:30 p.m. to 3:00 p.m. the trading volume was about ten times higher than the average trading volume for the same interval in the previous 30 days. The lesson learned from the graph of Figure 24 is that an high trading volume does not always imply improved liquidity: the high traded volume in the E-mini futures market resulted in a large order book imbalance which associates it at an extreme market volatility.

From 2:32 p.m. to 2:45 p.m. the Sell Algorithm sold 35.000 of the 75.000 E-mini contracts of the overall sell order, for a total revenue of almost $1.9 billions. The remaining 40.000 contracts were sold from 2:45 p.m. to 2:51 p.m. for a value of approximately $2,2 billions.

During the same period, after trading in the E-mini was reactivated at 2:45:33 p.m., the E-mini price progressively increased, reching the pre-crash levels by 3:08 p.m[187].

[187] Id.

3.4.5 Cross markets propagation of liquidity crisis and HFT activity

The liquidity crisis occurred in the E-mini market offered an opportunity for exploit cross markets arbitrage from cross-markets arbitrageurs, both opportunistic traders and HFTs; the most used financial instrument to implement the arbitrage strategy was the SPDR S&P 500 Trust ETF, also known as SPY ETF,the most traded ETF tracking the S&P 500 stock index. Since both the E-mini S&P 500 futures and the SPY ETF are derivative instruments of the same underlying, the S&P 500 stock index, their prices would move following a certain degree of correlation, maintaining constant the price ratio between the two derivatives when the underlying value changes. One SPY share price is equal to one tenth of the S&P 500 index; for instance, on May 6, 2010, when the S&P 500 was at 1.100 points, one SPY share's price was $110; thus, one E-mini futures contract is equal to 500 SPY shares[188].

The liquidity crisis in the E-mini futures and the abrupt decline of its price relatively to the SPY share price broke the aforementioned relation between the two instruments; this market inefficiency was exploited by cross market arbitrageurs: they buy the relatively undervalued instrument, the E-mini futures, and sell the relatively overvalued one, the SPY ETF on equity markets.

On May 6, 2010 the liquidity dinamic in the E-mini market was different from the one in the SPY market and many cross market arbitrageurs found that the price of the E-mini was relatively cheaper than the price of SPY during their price decline due to a larger selling pressure in the first market. Figure 25 shows the different dynamic in the buy side of the order books of the two instruments, where the buy side market depth is displayed as percentage of the morning average buy side market depth ; we can notice that by 2:40 p.m., a few minutes before the Flash Crash, the buy side market depth of the E-mini was reduced faster than the one of the SPY: the former was at less than 20% of the morning average, while the latter was at 75% of the morning average. Then, during the next five minutes, when the decline on the E-mini price was exacerbated by the Flash crash generated by the HFTs, the buy side depth significantly declined with respect to the one of the SPY: at the time when the CME "Stop Logic" functionality was activated, the buy side market depth of the E-mini was less than 1% of the morning average ($58 millions), while the buy side market depth of the SPY was declined to 25% of its morning average; thus, during this rapid decline arbitrage opportunities between the two instruments further increased[189].

The transmission of the liquidity crisis from futures market to equity market did not involve only the SPY ETF; cross market arbitrage strategies can be implemented also between the

[188] Id. 164.
[189] Id.

futures index and the basket of securities that constitute the underlying S&P 500 index and other ETFs. On May 6, 2010, the liquidity crisis transmission from the futures market caused 326 securities to be broken, for a total of 20.761 broken trades occurred on Nasdaq, NYSE Arca, BATS and various OTC markets: a trade is considered broken if it is executed at a price that excessively deviates from the last transaction price; stock markets consider these trades as erroneous trades because they do not reflect the fundamental value of the stock and consequently they cancel them[190]. A trade is considered broken when it execute against a stub quote; a stub quote is an offer to buy or sell at a limit price that is far away from the NBBO, an excessively high ask or low bid that are submitted by market makers with the intention to not execute it. Stub quote is an hodge used by designated market makers to avoid to trade when they are unwilling to trade or when the liquidity is exhausted while complaying with the affirmative obligation to continuously quote two-sided market for the securities with respect to which they are registered as market makers.

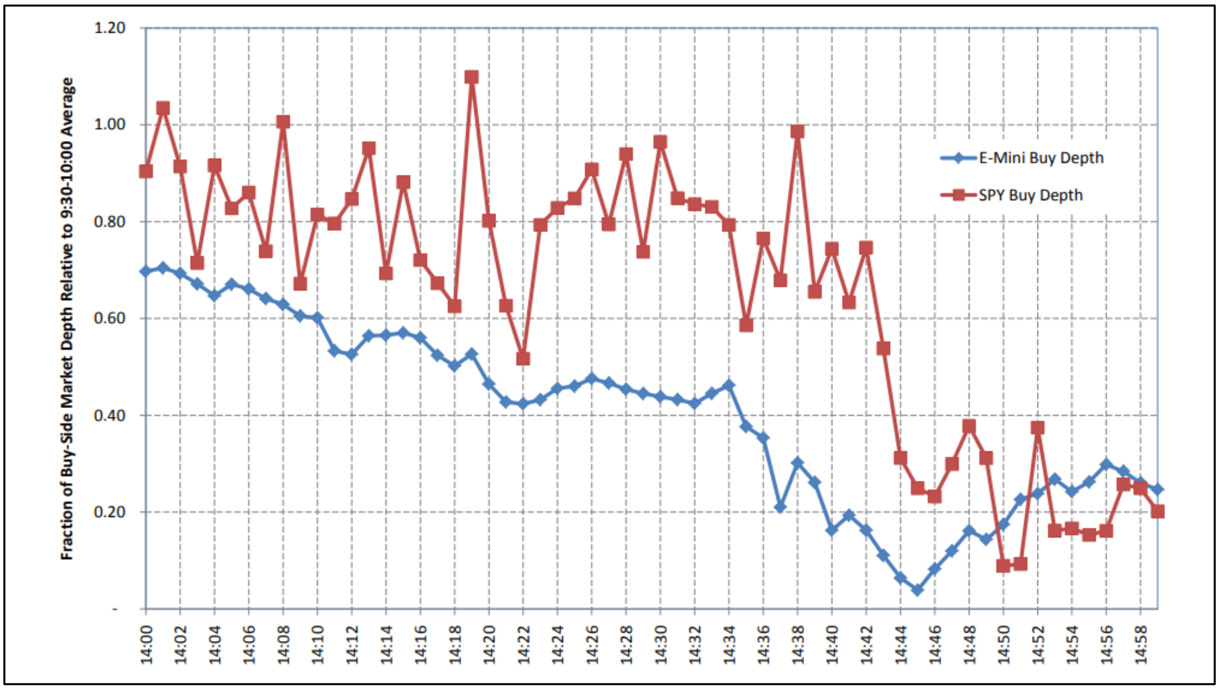


Figure 25: Buy-Side Market Depth for E-Mini and SPY
Source: Findings Regarding the Market Events of May 6, 2010

On May 6, 2010, between 2:40 p.m. to 3:00 p.m., the interval with the highest daily volatility and trading volume, 326 securities' price fell by 60% or more from their price at 2:40 p.m; for example Accenture plc declined in only 7 seconds, from 2:47:47 p.m. to 2:47:53 p.m., from $30 to $0.01 and Proctor & Gamble Co. declined from more than $60 at 2:40 p.m. to a low of

[190] Id.

$39.37 in less than 4 minutes[191].

The class of securities that were more affected by broken trades on May 6, 2010 were the ETFs: among the 326 securities that experienced broken trades on that day, 227 were ETFs. Figure 26 shows ETFs' timing of daily lows over one hour, from 2:00 p.m. to 3:00 p.m.: every point indicates the return of the ETF from the May 5 closing price to the lowest price reached on May 6, 2010. From the graph we can notice that ETFs' daily lows are mainly concentrated between 14:45 and 15, the period with peak volatility and trading volume; in particular, 160 ETFs experienced lows approximately 100% lower than the may 5 closing price, represented by the dense line along the -100% return on the y axis[192].

The activity on the ETFs was mainly due to HFT activity; the SEC-CFTC joint report analized the activity of 12 HFTs in securities listed on NYSE, NYSE Arca and Nasdaq on May 6, 2010 from the trading data reported by FINRA. With respect to ETFs, the trading volume of NYSE Arca listed securities (where ETFs are primarily listed) traded by HFTs during the price decline, from 2:43 p.m. to 2:46 p.m., increased by 254% with respect to the trading volume from 2:00 p.m. to 3:00 p.m., excluding the previous three minutes interval[193].

Figure 27 highlights that the significant increase in HFT trading volume on

[191] Id. 165.
[192] Id.
[193] Id. 117

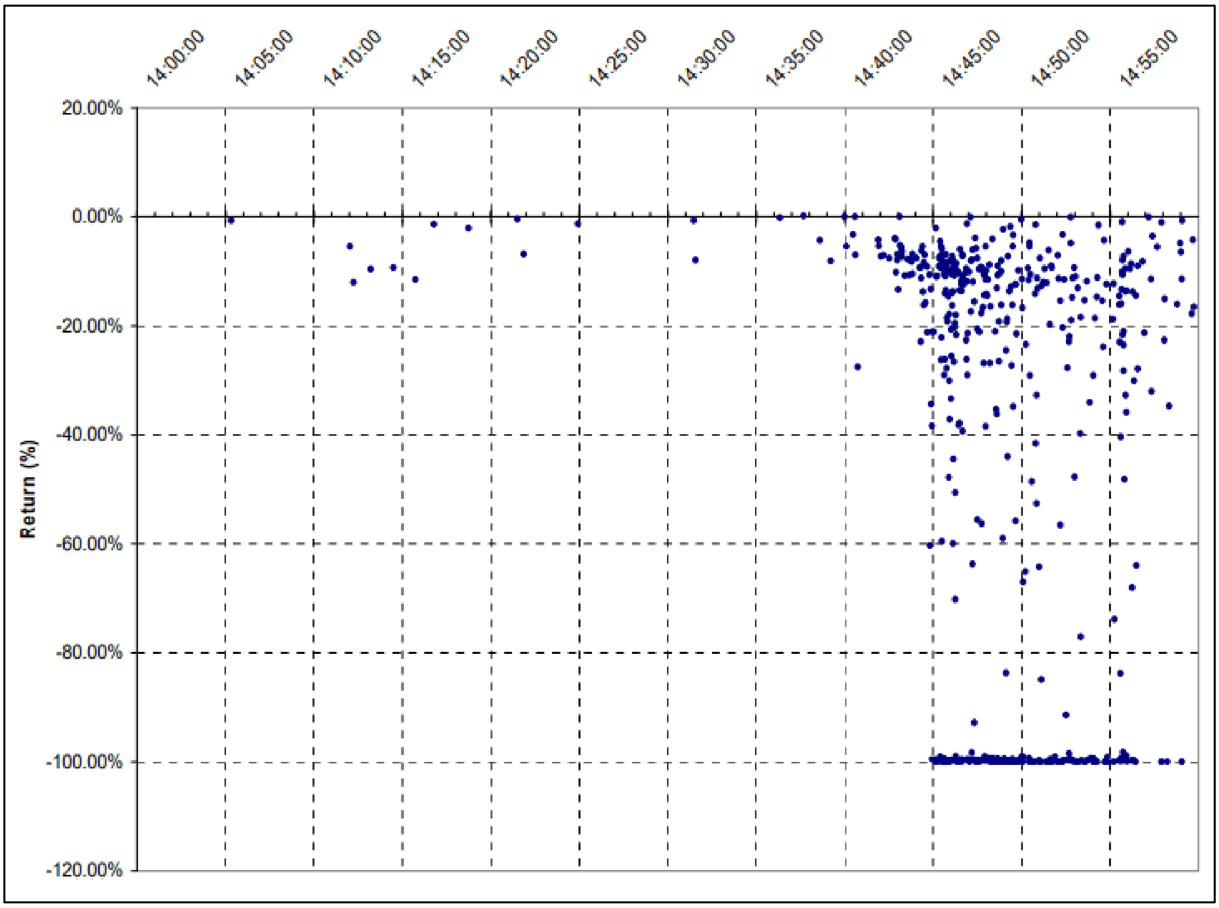


Figure 26: Timing of ETF Daily Lows, May 6, 2:00pm to 3:00 pm

Source: Preliminary Findings Regarding the Market Events of May 6, 2010

NYSE Arca is dominated by an escalation of selling activity that is consistent with the arbitrage strategy of buying the undervalued E-mini futures and selling the relatively overvalued ETFs; during the period of price decline ending 2:45 p.m. HFTs increased their aggressive selling more than any other trading categories in the three kinds of securities: from a net short position of $249 millions at 2:30 p.m. to a net short position of $1,158 billions at 2:45 p.m. This trading behavior is the same also for the HFTs trading activity on the NYSE and Nasdaq (where corporate stocks are primarily listed), even if with a much lower increase in HFT activity, respectively of 117% and 131% (Table 4). After this three time interval, HFT trading volume returned to the pre-2:43 p.m. level in all the markets.

In order to further examine the trading activity of HFTs during the market downward, the report also examines a data set obtained from the largest quoting markets during the May 6, 2010, from all the equities exchanges and Direct Edge. Table 5 shows the trading activity of 17 HFTs dividing it in aggressive and passive selling and buying activity; we can see that from 2:00 p.m. to 2:45 p.m. the trading activity of the 17 HFTs significantly increased to an intraday maximum of 50,3% of trading volume; in particular, we can notice that the increase in the HFTs' trading volume is mainly determined by aggressive selling activity, that increased more than the other categories in the period of market downward terminated at 2:45 p.m. (the aggressive selling trading volume increased to $9,3 billions). As we have previously seen, this increase in aggressive selling activity was mainly determined by cross-market arbitrage conducted by HFTs that simultaneously bought futures.

In the period within 2:46 p.m. to 3:00 p.m. the HFT activity dropped to 36,6% trading volume, that is consistent with what FINRA reported on the trading activity of 12 HFTs operating that day: After 2:45 p.m., when the indexes fell to their minimum, FINRA reported that 6 of the 12 HFTs reduced their trading activity, while 2 HFTs completely stopped their activity for the rest of the trading day. The last 4 HFTs significantly reduced their activity for short periods of time, from 1 minute (from 2:46 p.m. to 2:47 p.m.) to 21 minutes (from 2:57 p.m. to 3:18 p.m.).

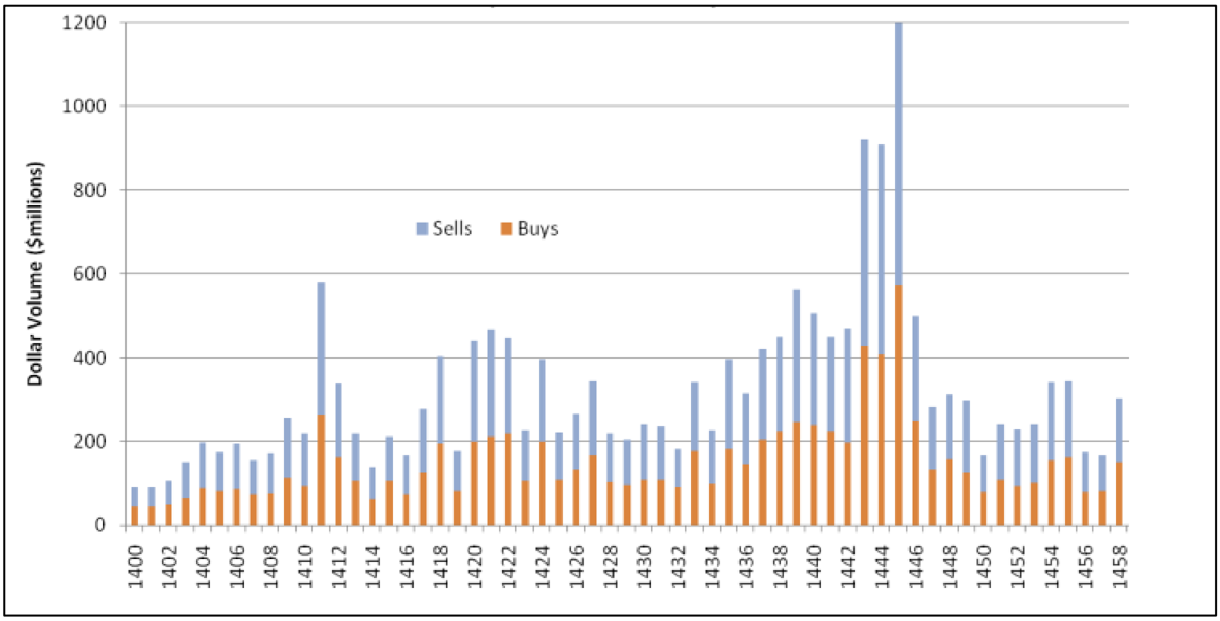


Figure 27: Dollar Volume of High-Frequency Traders for NYSE Arca-Listed Securities
Source: Findings Regarding the Market Events of May 6, 2010

Market	2:43 to 2:46 p.m. ($ Mil.)	2:00 to 3:00 p.m., ex 2:43 to 2:46 p.m. ($ Mil.)	Percent Change
NYSE	368.8	168.0	117%
NYSE Arca	1,011.7	285.6	254%
Nasdaq	310.9	134.7	131%

Table 4: HFT Trading activity
Source: Findings Regarding the Market Events of May 6, 2010

	HFT ($ Millions)							% of Total Market				
	Aggressive			Passive				Aggressive		Passive		
	Sell	Buy	Net	Sell	Buy	Net	Long/Short	Sell	Buy	Sell	Buy	Total
9:45 AM	2,674	2,904	230	3,044	2,723	-322	-92	39.2%	37.8%	40.5%	40.9%	42.2%
10:00 AM	2,449	2,447	-2	2,278	2,331	53	51	38.8%	41.3%	39.5%	37.8%	42.6%
10:15 AM	2,046	2,170	123	2,000	1,918	-82	42	39.8%	40.2%	37.9%	38.2%	41.5%
10:30 AM	2,141	2,128	-13	1,879	1,828	-51	-64	41.0%	44.1%	40.0%	36.0%	43.4%
10:45 AM	2,085	2,063	-22	1,789	1,790	1	-21	41.9%	45.1%	40.2%	36.9%	44.3%
11:00 AM	2,654	2,785	131	2,432	2,424	-9	122	40.4%	47.0%	42.1%	37.7%	45.3%
11:15 AM	2,667	2,728	61	2,443	2,396	-47	15	39.4%	47.1%	43.2%	36.2%	44.8%
11:30 AM	2,224	2,659	435	2,669	2,214	-454	-19	38.9%	40.8%	41.8%	39.6%	43.9%
11:45 AM	1,683	1,805	122	1,631	1,612	-19	103	38.0%	44.4%	41.2%	37.3%	43.3%
12:00 PM	2,316	2,695	379	2,549	2,274	-275	104	40.3%	45.5%	44.1%	40.6%	46.2%
12:15 PM	1,790	2,145	355	2,010	1,792	-218	137	41.2%	41.7%	40.0%	42.4%	44.6%
12:30 PM	1,390	1,422	32	1,276	1,230	-46	-14	41.0%	45.4%	42.0%	37.3%	44.5%
12:45 PM	1,324	1,339	15	1,115	1,136	20	35	43.2%	47.4%	40.8%	38.2%	45.8%
1:00 PM	1,624	1,720	96	1,560	1,437	-123	-27	42.8%	47.1%	44.0%	38.9%	46.2%
1:15 PM	1,642	1,434	-208	1,233	1,318	85	-123	42.9%	45.8%	40.6%	35.5%	44.4%
1:30 PM	2,294	2,425	131	2,269	2,139	-130	1	39.9%	46.6%	44.8%	38.1%	46.0%
1:45 PM	1,834	1,919	85	1,811	1,688	-123	-38	39.0%	45.5%	44.2%	36.8%	44.6%
2:00 PM	1,834	1,871	37	1,879	1,651	-228	-191	38.2%	41.8%	43.3%	35.4%	42.6%
2:15 PM	4,002	3,955	-47	3,739	3,517	-221	-268	41.2%	47.6%	46.4%	37.1%	46.0%
2:30 PM	5,786	5,814	28	5,571	5,294	-277	-249	44.6%	49.0%	48.2%	41.8%	49.3%
2:45 PM	9,302	7,959	-1,343	7,528	7,714	185	-1,158	47.1%	51.8%	50.2%	39.9%	50.3%
3:00 PM	5,748	5,071	-677	5,575	5,480	-95	-772	34.0%	32.9%	37.8%	33.5%	36.6%
3:15 PM	5,820	5,054	-765	5,515	5,428	-86	-852	46.3%	42.2%	47.3%	44.3%	47.0%
3:30 PM	5,220	4,732	-488	4,823	4,984	160	-328	43.2%	45.0%	47.0%	42.2%	46.0%
3:45 PM	4,763	4,547	-216	4,677	4,324	-353	-568	42.6%	41.0%	43.4%	39.8%	43.7%
4:00 PM	6,173	6,561	388	7,658	7,194	-465	-76	33.3%	37.8%	45.7%	40.1%	40.6%

Table 5: Dollar Volume of 17 High Frequency Trading Firms in Public Quoting Markets on May 6
Source: Findings Regarding the Market Events of May 6, 2010

3.5 Systemic risk

The May 6, 2010 Flash Crash has highlighted the risk for financial stability derived from poor designed algorithms and the exacerbation of the systemic risk by HFTs. Although the "Sell Algorithm" was intentionally executed with pre-specified instructions, the event has shed light on the vulnerability of algo trading to machine driven errors and biases, what Pereira (2020) calls cyber risks.[194]

Pereira identifies two main machine driven risks, that are the operational risk and the risk of inherent errors; the operational risk is associated with the strong reliance of algorithmic trading on automated trading, from the routing to the execution of the order, which makes them vulnerable to automated computer system malfunctions, such as software glitches, erroneous source code or disruptions in connectivity.[195] The risk of inherent errors arises from the fully disintermediation of algorithmic trading: they work with no or very limit human intervention, programmed in advance and autonomously managed by sophisticated computers; this high disintermediation makes more difficult for humans to intervene in real time to timely solve potential technological malfunctions, in particular if the trading strategy occurs at high-frequency, with the consequence that many damages can be done before humans can intervene to stop the malfunctioning algorithm.[196] Furthermore, since algorithmic trading strategies are not as varied as those implemented by human traders (they tend to be designed to react to the same input in correlated ways), the high correlation between them increases the market systemic risk, that is the risk that a malfunction in an algorithm implemented by a single trader can largely widespread across traders and markets, as we have seen in the cross-markets liquidity crisis propagation in section 3.4.5.

One of the most famous HFT default caused by malfunctioning in its execution algorithm was the August 1, 2012 Knight Capital trading glitch. On that day, Knight Capital Group, Inc., a large HFT group engaged in market making, experienced a malfunctioning in the code sequence of one of its automated order routers, which resulted in the HFT firm sending 4 millions orders into the market, which ultimately led to the trading of 397 millions of shares in 45 minutes.[197] During the abnormal trading activity generated by the broken algorithm, Knight Capital lost approximately $440 millions, three times its annual earning, and the effect widespread across stocks, with 150 stocks experiencing significant price movements and across markets, with significant variations in trading activity not only on the NYSE, where the erroneous orders were sent, but also on Nasdaq and NYSE Arca[198].

[194] Clara Martins Pereira, 2020, Unregulated Algorithmic Trading: Testing the Boundaries of the European Union Algorithmic Trading Regime, Journal of Financial Regulation, Vol. 6, 270-305.

[195] Id.

[196] Id.

[197] SEC, Press Release, SEC charges Knight Capital with violations of Market Access Rule (Oct. 16, 2013), available at https://www.sec.gov/news/press-release/2013-222.

[198] Charles R.Korsmo, 2013, "High-Frequency Trading: A Regulatory Strategy", U. Rich. L. Rev., Vol. 48,523-609.

3.6 Market efficiency

Market efficiency refers to the degree to which market prices reflect all the available informations. According to Fama (1970), a market is efficient when prices always fully reflect all the available informations, which implies that security's price is the most accurate estimate of the security's fundamental value; thus, market participants cannot make profits on new informations because once they arrive they are immediately incorporated in the price.[199] Efficiency in processing informations improves capital markets' functioning, whose main purpose is to efficiently allocate investors' capital: the larger is the quantity of informations that prices reflect, the easier is for market participants to understand the true value of a security and to decide the investment project that has the capacity to most profitably use their capital. Therefore, informational efficiency implies allocative efficiency. Furthermore, once the capital is allocated, the expressive functionality of prices can function as a monitoring and disciplinary device for capital. If stocks'prices work efficiently, they will reflect the rational traders' expectations on the performance of the firms issuing them, which mainly depends on the evaluation of how their management works; thus, efficient prices offer a signal of good or bad corporate management, allowing shareholders to monitor corporate managers' performance and also helping managers understand how they and the firm are performing through a feedback mechanism between them and the market[200]. The monitoring role of prices allows shareholders to incentivize managers to pursue company's interest by tying their payments to shareholders return, for example through the use of stock options. In addition to monitoring, price signals can also trigger a disciplinary mechanism for corporate management; with reference to the market for corporate control, when the price of a security is low because of bad management and not for a slump in the market or other systemic reasons, takeover specialists will exert discipline through an hostile takeover with the aim to run the company in a more profitably way[201].

Informations are incorporated into prices when traders trade on their estimates of what a security is worh: every time traders buy or sell a security on the basis of the informations they have collected on the security's value, the price incorporates this piece of information changing and informing investors of the new information. The mechanism through which capital markets actually reflect information can be explained by the seminal work of Gilson and Kraakman (1984) which identified four kind of traders that interacting one another make markets efficient:

---

[199] E. Fama, 1970, Efficient Capital markets: a review of theory and empirical work, The Journal of Finance, Vol 25, No. 2, 383-417.

[200] Yesha Yadav, 2015, How Algorithmic Trading undermines efficiency in capital markets, Vanderbilt Law Review, Vol. 68 (6), 1607-1671.

[201] Id.

universally informed traders, professionally informed traders, derivatively informed traders and uninformed traders[202]. Universally informed traders trade on the base of informations that are simultaneously available to all the investors, for instance, the informations that are already reflected in the price; in other words the price itself is an information simultaneously available to all the investors that incorporate all the past informations. In the 1980s, economists used as an example of information immediately available to all the investors the U.S. presidential election results[203].

Professionally informed traders, also called well-informed or fundamental traders, trade on the basis of publicly available informations that can be efficiently used only by sophisticated investors, such as market analysts, industry experts and professional asset managers. It is not necessary that all the investors are able to use this kind of informations: it is sufficient that a relatively large group of investors use these informations to incorporate them into the price; furthermore, sophisticated traders are usually institutional traders, whose size and skills are able to consistently move prices[204]. Informed traders invest considerably resources to obtain as much informations as possible that can give them an edge with respect to other market participants; if on one hand less-informed or uninformed traders suffer losses trading with this kind of traders, on the other hand informed traders make markets more efficient: their private interest generates public gains[205]. Furthermore, informed traders' fundamental research can spot price discrepancies between similar and correlated financial instruments and correct them through arbitrage strategies.

Derivatively informed traders, also called order anticipators, are parasitic traders who profit at the expense of informed traders; differently from informed traders, they do not trade on the basis of fundamental research, but on the ability to predict how informed traders will trade by observing their trading behaviour. This group of traders reduce informed traders' returns but, at the same time, accelerate the price discovery process and make markets even more efficient[206].

Finally, uninformed traders trade without knowing whether financial instruments are undervalued or overvalued because either they cannot form reliable opinions about values or they choose not to. Uninformed traders generate what Fisher Blake first called noise trading; he stated that in financial markets traders do not trade only on the basis of informations, but

---

202 Ronald Gilson & Reinier R. Kraakman, 1984, The Mechanisms of Market Efficiency, Virginia Law Review, Vol. 70, No. 4, 549-644.
203 A.Lupoi, 2020, La struttura del mercato ed i riflessi giuridici, CEDAM.
204 Id.
205 Id. 200
206 Id. 79.

also on what they believe to be informations, that he called noise[207]. Consequently, even if uninformed traders provide liquidity to the market, they move prices away from their fundamental value.

3.6.1 HFT effects on informational efficiency

HFTs offer many benefits to market efficiency; their capacity to quickly react to and process new informations through the use of low-latency infrastructures and highly sophisticated automated algorithms makes prices more responsive than ever before to incoming news. HFTs' technological advantage does not only make markets better able to quickly reflect available informations, but it also incorporate in prices a richer reserve of informations: beyond conventional data sources such as macroeconomic news or prices, they have the computational power to collect and process data from a diffuse range of resources, such as social media database like Twitter or Facebook that are used to spot prevailing sentiments and trends[208]. Finally HFTs' ultra-fast trading infrastructure enables them to quickly spot and correct price discrepancies between similar and correlated instruments, even across different markets, enhancing the effectiveness of arbitrage; thus, HFT activity quickly corrects price inefficiencies, enabling prices to reflect more accurately their fundamental values. One of the most important finding on the role played by HFTs on price discovery process was provided by Brogaard , Hendershott and Riordan (2013); in their working paper, the authors found positive effects of HFT activity on market informational efficiency: they trade against transitory price movements and in the direction of permanent price movements.[209] Transitory price movements, also called pricing errors or noise, are determined by temporary liquidity imbalances due to uninformed traders; on the contrary permanent price movements are due to the incorporation in the price of informations about the asset's fundamental. The authors investigate the correlation between HFT activity and the average market return of a sample of 120 stocks listed on Nasdaq and NYSE from 2008 to 2009, dividing the activity of the Nasdaq flagged 26 HFT firms in liquidity demanding (HFT^D) and liquidity providing activity (HFT^S)[210]. Figure 28 shows the correlation between HFT liquidity demanding activity and non HFT liquidity demanding activity and future and past market returns. HFT^D is positively correlated with future market returns, even if this correlation quickly dies (the correlation

[207] Blake F., 1986, Noise, The Journal of Finance, Vol. 41 (3), 529-543.
[208] Id. 200.
[209] J. Brogaard, T. Hendershott, R. Riordan, 2013, High Frequency Trading and price discovery, European Central Bank Working Paper Series No. 1602.
[210] HFT and non-HFT trading variables are expressed in terms of order flow (net trading): buy volume minus sell volume.

coefficient goes to 0 after three seconds); this finding is consistent with the capacity of HFTs to predict future price movements and thus to trade in this direction (they buy before prices go up and sell before they go down) but this correlation lasts only a few seconds because informations incorporated by HFTs are short-lived[211]. On the left side of the graph we see a negative correlation between past market returns and HFT^D, which is consistent with the capacity of HFTs to correct transitory price movements, following a contrarian trading strategy (they buy after prices have decreased and sell after prices have increased)[212]. Non-HFT liquidity demanding traders $nHFT^D$ also predict future price movements but their trading volume is positively correlated with past market returns, which means that they tend to follow market trends[213]. Figure 29 shows the correlation between HFT^S and non-HFT liquidity providers' activity ($nHFT^S$) and past and future market returns; both the variables are negatively correlated with future market returns, even if the correlation coefficient for HFT^S is lower: this is consistent with the fact that market makers are usually adversely selected (prices go up after they have sold and go down after they have bought), with HFT market makers being less adversely selected. Finally, $nHFT^S$ and HFT^S are respectively positively and negatively correlated with past market returns: $nHFT^S$ tend to follow momentum strategy while HFT^S contrarian trading helps to correct temporary noises in the market[214].

The aforementioned results hold also when the authors decompose the price movement in temporary price movements and transitory price movements; in conclusion, HFTs benefit markets because they eliminate transitory pricing errors reducing the noise in the price discover process and accelerate and increase the incorporation of new informations in the prices.

[211] Id. 209.
[212] Id.
[213] Id.
[214] Id.

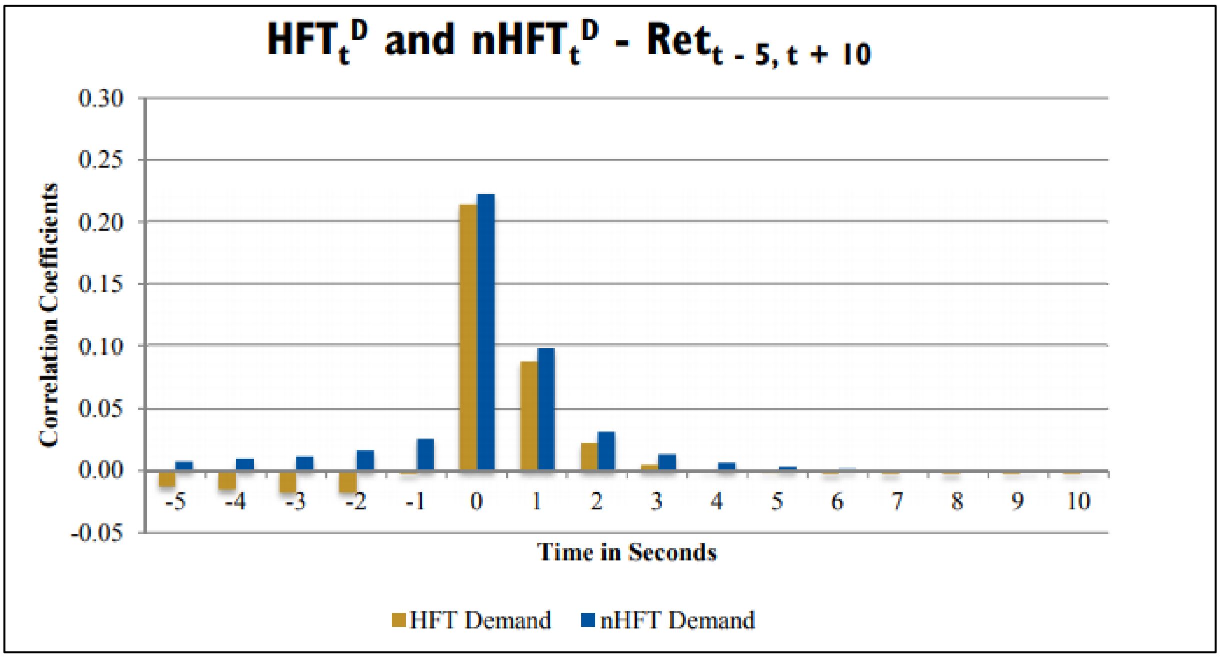


Figure 28: Correlation of returns with HFT and nHFT liquidity demand

Source: J. Brogaard, T. Hendershott, R. Riordan (2013)

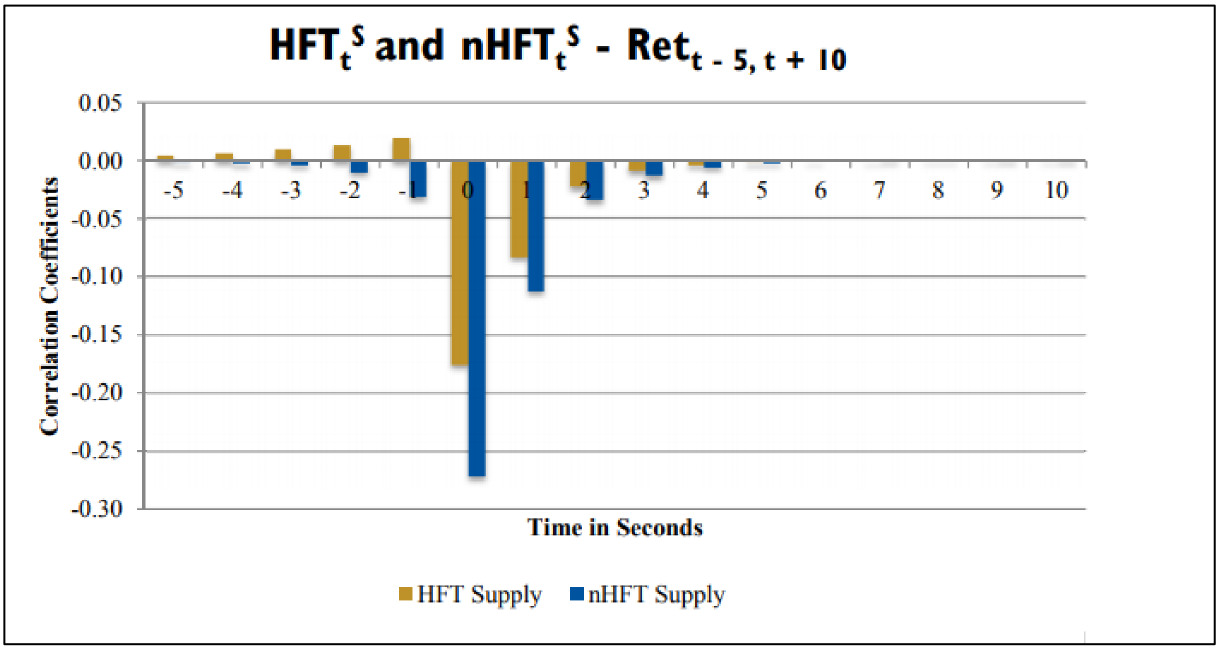


Figure 29: Correlation of returns with HFT and nHFT liquidity supply

Source: J. Brogaard, T. Hendershott, R. Riordan (2013)

3.7 Structural insider dealing

In the previous section we have seen how the use of co-location services, direct market feeds and automated algorithms by HFTs have contributed to market efficiency, improving and accelerating the way informations are incorporated into the prices; despite these benfits, criticism of HFT point out how the use of low-latency infrastructures has created two sided market, increasing the asymmetric information between HFTs and non-HFTs. Yadav (2016) argues that HFTs' use of low-latency infrastructures has created in financial markets the same uneven playing field generated by insider trading; for this reason he calls the practice of HFTs of using ultra-fast infrastructures structural insider dealing[215]. As insider traders, HFTs become aware of informations that are likely to affect securities' prices before they become publicly available and thanks to this edge they are able to better perform than everyone else in the market. However, structural insider dealing does not fall within the scope of illegal insider dealing because the way through which HFTs collect the informations does not breach any fiduciary duty, as on the contrary holds for corporate insiders who use their privileged access to corporate's informations to profit at the expense of the company. HFT's informational advantage is a structural advantage: as we have seen in section 2.3.1, through the use of co-located services and direct market feeds, HFTs have access to new market informations before they are reported to consolidated market data subscribers; thus, HFTs see the orders of other market participants after they actually are submitted and made public. Furthermore, direct market feeds transmit a richer quantity of informations with respect to consolidated market data: while the SIP only collects the BBO from the trading venues of the National Market System to provide investors the NBBO, direct market feeds are customized packaged quantities of market data that allow subscribers to have access to more detailed informations of other customers' orders, such as order cancellations, modifications and executions, in order to make the best trading decision[216]. Figure 30 schematically shows the information flow within an exchange with HFTs. At time 0 the price of a stock is equal to \$100; an incoming information that is likely to affect the stock price is routed simultaneously to the SIP and the direct feed; thanks to the market feed co-located server, HFT receives at time 1 the new information before all the other market participants, and within a few microseconds trades on the new information. The HFT's price impact increases the price of the stock to \$101 at time 2, before the new information has reached other investors with slower direct market feed at time 3 and SIP subscribers at time 4. Thus, at time 3 and 4 slower traders will trade on a stale information and

[215] Yesha Yadav, 2016, Structural insider dealing, UCLA Law Review, Vol. 60, 968-1033.
[216] Id.

will be adversely selected by the HFT, which will close its position at a gain[217].

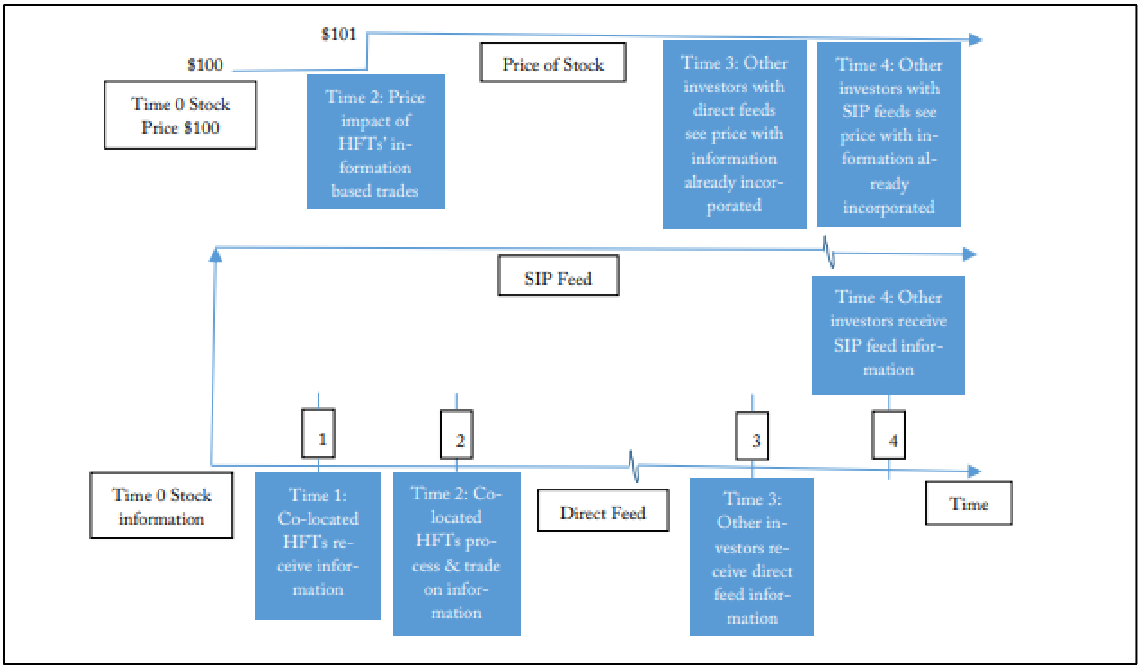


Figure 30: Structural Insider Dealing

Souce: Yadav (2016)

The structural advantage of HFTs is detrimental for non-HFTs' revenue: if HFTs can systematically see market data before all public investors, they can anticipate how large informed traders will trade and trading in the same direction ahead of them, raising the price that large traders will pay, or, through slow-market arbitrage strategies, even fill small orders at inferior prices (section 2.3.1). This systematic advantage disincentives traders that are systematically hurted by HFTs to trade: they will trade less, investing lower capital in fundamental research or they leave the market. The withdrawal of liquidity by investors, in particular those informed, impairments market quality; investing less in fundamental research reduces the informational power of prices, while the reduction of trading volume drains markets of their power to allocate capitals, leaving public companies with lower funding opportunities[218]. Even if actually no large traders exit the market, they leave light stock exchanges, trading in dark pools, where it is more difficult for HFTs to detect their order flows. The migration of investors to dark pools makes trading less trasparent and more difficult and costly for traders to interpret the signals of their trading, reducing the informational efficiency of markets.

217 Id.

218 Id.

Exchanges have no incentives to solve the asymmetric information between HFTs and non-HFT by filling the informational gap between direct market feeds and SIP: they can earn higher revenues by traders willing to pay for having a more timely and detailed access to market informations. The ability of exchanges to commoditize and sell informations has undermined market quality: the increased cost of procuring informations through exchanges' specific products has increased the market entry cost and reduced investments in fundamental research[219].

Asymmetric information arises also when HFT leverages flash trading facilities; as we have seen in section 2.3.2 , flash facilities give HFT the opportunity to execute an incoming order at the NBBO or better within window of a few milliseconds before routing it to the exchange quoting the NBBO. Thus, trading facilities give HFTs an informational advantage with respect to non-HFT that HFT can use to exploit two opportunities: if the HFT can profitably fill the flash order, it will follow one of the three trading patterns described in section 2.3.2 , while if it decides to not fill the flash order, knowing in advance the trading venue with the NBBO where the order will be routed, it can cancel limit order previously posted on that exchange to avoid adverse selection losses generated by the liquidity pressure of the incoming order[220].

[219] Id.

[220] Id. 120.

4 US AND EU REGULATION

4.1 Oversight expansion on trading activity

The risks of algorithmic trading, in particular HFT activity, for financial market stability and integrity highlighted the need of a securities market regulation that was able to keep up with the technological innovations that was drastically changing the capital market landscape. In the wave of the 2010 Flash Crash and other flash crashes occurred during the years and in the face of high-speed backed manipulative trading strategies, US and EU regulators tried to enhance capital markets oversight and to restore investor confidence imposing more specific registration and disclosure requirements for market participants engaging in algorithmic and HF trading.

4.1.1 Large Trader Reporting Rule

In order to restore investor confidence in capital markets, securities market regulators must be able to detect manipulative trading strategies by amplifying their surveillance power on HFT strategies; in 2011 SEC passed the "Large trader reporting rule", which requires large traders to comply with specific registration and reporting requirements. Large traders are those whose trading activity in NMS securities exceeds one of the following "identifying activity level": either during a calendar day transactions in NMS securities has exceeded 2 millions of shares or shares for a fair value of $20 millions or during a calendar month 20 millions of shares or shares for a fair value of $200 millions.[221] Once exceeded one of these thresholds, traders have to register with the SEC as a large trader in order to receive a large trader identification number that it has to communicate to all the registered broker-dealers that trade on its behalf; broker-dealers in turn have to recordkeep all requested informations on transactions carried out by the large trader and report them to the SEC upon request[222]. The Large trader reporting rule allows US market regulators to monitor the trading activity of large traders, and in particular the activity of HFT, making easier for the SEC to reconstruct capital markets' dynamic and link HFT trading patterns to disruptive market events.

[221] 17 CFR § 240.13h-l (a)(7)(i)(ii).

[222] 17 CFR § 240.13h-l (b)(1), 17 CFR § 240.13h-l (d)(1): among the requested informations broker-dealers have to report there are the identifying symbol of the traded security, number of shares traded in each reported transaction, whether each transaction was a purchase, sale, or short sale, the transaction price, the time that the transaction was executed.

4.1.2 FINRA Registration for High-Frequency Securities Traders

In 2015 the SEC proposed to amend Rule 15b9-1 in order to enlarge its monitoring power on broker-dealers engaging in cross-market proprietary trading on off-exchange (ATSs or OTC), of which HFTs represent a significant proportion. Under section 15(8)(b) of the Security and Exchange Act, a broker-dealer must be registered with a national securities association (the only existing national securities association is FINRA) unless it trades only on the national securities exchange of which it is registered as a member[223]; Rule 15b9-1 provides excemptions to the previous provision, allowing broker-dealers to not register with FINRA if it is a member of a national securities exchange, carries no customer account and the gross income from transactions executed out of the exchange with which it is registered does not exceed $1000, the so called "minimis allowance"[224]. Furthermore, the "minimis allowance" does not include proprietary trading, but only transactions carried out on behalf of customer accounts, the so called proprietaty trading exception; consequently, the provision allows broker-dealer to trade on their own account off-exchange without being monitored by FINRA, increasing the risk that they engage in disruptive practices, also at high-frequency, that can impair capital market integrity and stability.

The amendment to rule 15b9-1 proposed to eliminate the proprietary trading exception, requiring that broker-dealers must effect transactions only on the national exchange with which they are registered to be exempt from FINRA registration; the only exceptions would regard off-exchange trading for hedging risks of their exchange transactions and off-exchange trades executed to comply with the order protection rule (Rule 611) of Regulation NMS.[225] The registration with FINRA can cause some HFTs to leave off-exchange trading or to reduce their activity, given the increased cost of trading under FINRA oversight; even if off-exchange trading reduce bid-ask spread and increase market depth, an increasing off-exchange trading activity can cause prices on lit exchanges to no longer reflect the security's fundamental.[226]

[223] 15 U.S.C. §78, et seq.

[224] 17 CFR § 240.15b9-1; the exemption was initially designed to allow exchange members to conduct hedging or other off-exchange activity ancillary to their exchange activity.

[225] Press Release, SEC, SEC Proposes Rule to Require Broker-Dealers Active in Off-Exchange Market to Become Members of National Securities Association (March 25, 2015), https://www.sec.gov/news/pressrelease/2015-48.html.

[226] Michael Morelli, 2017, Implementing High Frequency Trading Regulation: A Critical Analysis of Current Reforms, Michigan Business & Entrepreneurial Law Review, Vol.6 (2), 201-229.

4.1.3 CAT and MIDAS

A further step to a more comprehensive registration regime was the adoption in 2012 by the SEC of Rule 613, which required national securities exchanges and FINRA to implement the Consolidated Audit Trail (CAT), a national market system data management plan to track the life cycle of all orders and trades[227]. Although exchanges report trades data to Security Information Processors, there are no nationwide audit trail keeping record of all order activity, including order cancellations (trade data particularly important for detecting HFT). The creation of a central repository of trading data not only help regulators to monitor and reconstruct trading activity, but CAT can enable private parties to better establish the causation link betwen HFT disruptive practices and market events and thus intent in class action claims.[228]
In 2013, SEC established the Market Information Data Analytics System (MIDAS) as the agency official trade monitoring system. Every day MIDAS collects and processes data on listed stocks, exchange-trade products, equity options and futures contracts from security information processors and exchanges' proprietary feeds, time-stamped to the microsecond; in particular, MIDAS collects posted orders and quotes on national exchanges, order cancellations and modifications, exchange and off-exchange trade executions.[229] MIDAS allows the SEC to have access to detailed informations of the order book in near real-time, enabling it to monitor market activities on a continuous basis.

4.1.4 Algorithmic trading notification requirements

On November 24, 2015, the CFTC approved Regulation Automated Trading (Regulation AT) in reponse to the growth of Algorithmic Trading Systems (ATSs), of which HFT is a dominant fraction, with the aim to provide "a series of risk controls, transparency measures, and other safeguards to enhance the U.S. regulatory regime for automated trading".[230]
Under Reg AT, a market participant engaged in algorithmic trading has to notify to its clearing members and the designated contract market (DCM)[231] on which it trades that it will engaged in algorithmic trading before submitting an order to a DCM.[232]
EU financial markets are regulated by Directive 2014/65/EU of the European Parliament and

---

[227] 17 CFR § 242.613
[228] Id. 226.
[229] SEC, MIDAS: Market Information Data Analytics System, https://www.sec.gov/marketstructure/midas-system.
[230] Press Release, CFTC, CFTC Unanimously Approves Proposed Rule on Automated Trading (Nov. 24, 2015), https://cftc.gov/PressRoom/PressReleases/7283-15.
[231] See Trading Organization, CFTC, https://www.cftc.gov/IndustryOversight/TradingOrganizations/index.htm; Designated contract markets (DCMs) are exchanges that may list for trading futures or option contracts.
[232] Regulation AT §1.80(d).

of the Council of 15 May 2014, known as the Market In Financial Instruments Directive II (MIFID II), in force since January 3, 2018, which replaced the former Directive, MIFID I; the new regulation addresses algorithmic trading with provisions specifically tailored for this breed of traders. With respect to algorithmic trading notification, MIFID II requires that all investment firms that want to engage in algorithmic trading have to notify this to their National Competent Authorities (NCAs) and to the trading venue where they trade as a memeber or participant.[233]

4.1.5 Source code disclosure

Financial markets suffer by an asymmetric information arisen from a lack of transparency on how algorithmic models work; the absence of a regulatory scrutiny on their functioning cannot alert financial markets regulators on mulfanctioning or "rogue", poorly designed algorithms before they lead to disruptive market events, as happened with May 6, 2010 Flash Crash or Knight Capital default. Furthermore, the absence of regulatory monitoring on algorithmic models can incentivize traders to heighten the complexity of it, increasing the probability of glitches[234]. In order to fill this informational gap, Regulation AT requires algorithmic traders to "Maintain a source code repository to manage source access, persistence, copies of all code used in the production environment, and changes to such code; […] Each AT Person shall keep such source code repository, and make it available for inspection, in accordance with § 1.31"[235] Similarly to Regulation AT, MIFID II requires significant disclosure for algorithmic traders, stating that "The competent authority of the home Member State of the investment firm may require the investment firm to provide, on a regular or ad-hoc basis, a description of the nature of its algorithmic trading strategies, details of the trading parameters or limits to which the system is subject […]".[236] With unrestricted access to the source code of algorithmic traders, financial regulators can better detect poor designed algorithm or algorithm intentionally designed to implement illegal HFT strategies, such as spoofing or quote stuffing. However, the CFTC source code repository provision has been criticized by the industry; the algorithmic source code represents a fundamental asset, the "lifeblood" of algorithmic traders' business, that should be protected by laws safeguarding industry trade secrets: "While we think that the CFTC's goal is perfectly reasonable, it's inconceivable that any firm should be expected to

233 Article 17(2) Directive 2014/65/EU.
234 Id. 200.
235 Regulation AT, §1.81(a)(6). See 17 CFR § 1.31(d)(1) which requires that regulatory records shall be open to inspection by SEC or United Stated Department of Justice.
236 Article 17(2) Directive 2014/65/EU.

leave its intellectual property on the doorstep of the government," said Bill Harts, chief executive of Modern Markets Initiative.[237]

4.2 Requirements for internal system controls

Capital market regulators strenghtened pre-trade and internal risk controls on traders engaging in algorithmic trading, in particular HFT, with the aim to reduce the systemic risk genereted by "rogue" algorithms, in particular during stressed market conditions. Under section 1.81 of Regulation AT, CFTC establishes that algorithmic traders have to test the algorithmic code and related trading strategy before its deployment on the relevant DCM; through the use of a development environment, isolated from the production trading one, algorithmic traders have to test the resiliency of their trading strategies. Algorithmic traders have to conduct regular back-testing of algorithmic trading on actual historical data to test ex-post how the algorithmic trading system would have performed during specific periods, in particular turbulent times, and thus to see how the algorithm would perform under future similar circumstances; therefore, they have to regularly conduct stress tests of algorithmic trading systems to verify their resiliency under a variety of market situations.[238] Section 1.80 of Regulation AT provides specific pre-trade risk controls for algorithmic traders, which include caps on message frequency and trade execution frequency over a specific interval, order price parameters and maximum order size and Order Cancellation Systems to disengage algorithmic trading or cancel orders under certain market conditions; these pre-trade risk controls prevent HFT to flood the market with excessive liquidity, adressing ghost liquidity issues, and enable algorithmic traders to correct the sending of erroneous orders.[239]

EU regulators followed a similar regulatory pattern; MIFID II specifically requires investment firms engaging in algorithmic trading to adopt "effective systems and risk controls suitable to the business it operates to ensure that its trading systems are resilient and have sufficient capacity, are subject to appropriate trading thresholds and limits and prevent the sending of erroneous orders or the systems otherwise functioning in a way that may create or contribute to a disorderly market".[240] In particular, the Directive charges ESMA to provide draft regulatory technical standards to specify the organizational requirements of investment firms engaged in

[237] See Gregory Meyer, Industry Criticizes CFTC's Plans For New Automated Trading Rules, FIN. TIMES (Mar. 16, 2016), https://next.ft.com/content/d6558728-ebba-11e5-9fca-fb0f946fd1f0; Gregory Meyer and Philip Stafford, US regulators propose powers to scrutinise algo traders' source code, FIN. TIMES (Dec. 1, 2015), https://www.ft.com/content/137f81bc-944f-11e5-b190-291e94b77c8f.

[238] Regulation AT §1.81(a)

[239] Regulation AT §1.80(a) and (b).

[240] Article 17(1) Directive 2014/65/EU.

algorithmic trading.[241] Under Commission Delegated Regulation (EU) 2017/589 of July 19 2016, known as Regulatory Technical Standard 6 (RTS 6), ESMA provides organizational requirements for algorithmic traders; as under Regulation AT, RTS 6 requires algorithmic traders to test the algorithm before any depoloyment or update on a testing environment separated by the production one (the trading venue where the trader want to implement their algorithmic trading strategy), following a methodology that ensures the algorithmic system and strategy do not behave in an unintended manner, as for market manipulative purposes and do not contribute to disorderly markets.[242] Investment firms have to conduct a stress test on an annual basis of its algorithmic trading system to verify their resiliency against increased order flows or market stresses; in particular, the provision indicates two stress tests that at least the investment firms must conduct: an high message volume test, simulated multiplying by two the highest number of messages received and sent by the investment firm in the previous six months, and an high trade volume test, multiplying by two the investment firm's highest trading volume in the previous six months.[243] Furthermore, MIFID II requires trading venues to provide to market participants engaged in algorithmic trading testing environments to facilitate the aforementioned testing.[244] Commission Delegated Regulation (EU) 2017/584 of 14 July 2016 (RTS 7), which supplements MIFID II providing organizational requirement of trading venues, establishes that trading venues have to provide to their members access to a testing environment which consists of "simulation facilities which reproduce as realistically as possible the production environment, including disorderly trading conditions, and which provide the functionalities, protocols and structure that allow members to test a range of scenarios that they consider relevant to their activity"[245].

As for US regulation, MIFID II requires that investment firms engaged in algorithmic traders have in place pre-trade controls, which include price collars, a mechanism that automatically block or cancel orders that do not meet specified price parameters, maximum order values, maximum order volume and maximum message limits (submission, modification and cancellation) to prevent excessive HFT activity.[246] Furthermore, algorithmic traders have to adopt a "kill functionality", which allows them to timely cancel, as an emergence measure, their unexecuted orders, in particular to prevent erroneous orders from "fat finger errors" or computer

[241] Article 17(7) Directive 2014/65/EU.
[242] Articles 5 and 7 Delegated Regulation 2017/589/EU.
[243] Article 10 Delegated Regulation 2017/589/EU.
[244] Article 48(6) Directive 2014/65/EU.
[245] Article (10)(a) Delegated Regulation 2017/584/EU.
[246] Article 15(1) Delegated Regulation 2017/589/EU.

glithes.[247]

4.3 Circuit Breakers

Beside the internal risk controls of algorithmic trading systems that we have seen in the previous section, also trading venues have implemented control systems to prevent disorderly markets. Since the 1987 stock market crash, the so called "Black Monday", SEC adopted a Market Wide Circuit Breaker (MWCB) to address excessive market volatility. In force from October 1988, the MWCB triggered a cross-market trading halt if the daily variation of the Down Jones Industrial Average (DJIA) exceeded pre-determined Circuit Breaker thresholds, in terms of percentage of a reference price. SEC calculated quaterly the reference price as the average of the DJIA closing values over the previous month; the CB thresholds were 10% (level 1), 20% (level 2) and 30% (level 3) daily decline from the reference price, which triggered different cross-market trading halts depending on the time of the trading day[248]:

- The halt for a level 1 decline was one hour if it occured before 2 p.m., 30 minutes between 2 p.m. and 2:30 p.m. and no halts after 2:30 p.m;
- The halt for a level 2 decline was two hours if it occurred before 1 p.m., one hour between 1 p.m. and 2 p.m., and market closure for the rest of the trading day after 2 p.m.;
- If a level 3 decline occurred, trading is closed for the rest of the day regardless of the time it occurred.[249]

The first response of exchanges and FINRA to the events of May 6, 2010 Flash Crash was the implementation of the Single-stock Circuit Breaker pilot program (SSCB). The SSCB was implemented progressively including an ever larger group of stocks: on June 2010 the SEC approved the pilot program for stocks included in the S&P 500 index, then it extended the program to Russel 1000 index stocks and on June 13, 2011 all the remaining NMS stocks were included; SSCB was active from 9:45 a.m. to 3:35 p.m. and it halted trading for at least five minutes in a stock that moved up or down by 10% within five minutes[250].

On June 2012, the SEC approved two proposals submitted by the FINRA and the national securities exchanges to update the MWCB and the SSCB. The new MWCB replaced the DJIA

---

[247] Art. 12(1) Delegated Regulation 2017/589/EU.

[248] Sullivan & Cromwell LLP, 2012, SEC Order Approves Proposals to Address Extraordinary Volatility in Individual Stocks and Broader U.S. Stock Market.

[249] Avanidhar Subrahmanyam, 2012, Stock Market Circuit Breakers, Government Office for Science.

[250] Claudia E. Moise, Paca Flaherty, 2017, Limit Up-Limit Down Pilot Plan and Associated Events, U.S Security and Exchange Commission, white papers.

index with the S&P 500 index as reference stock index for cross-market trading pause in order to take into account a broader index in terms of market capitalization (the DJIA index is computed only on 30 stocks); the CB thresholds were reduced to 7% (level 1), 13% (level 2) and 20% (level 3) with respect to the previous day's closing price (trigger thresholds were reculculated daily and not quaterly) and the temporal structure of CBs was simplified to two relevant trigger periods: a level 1 and 2 market decline before 3:25 p.m. triggered a trading halt of 15 minutes, while no trading pauses were triggered if they occurred after 3:25 p.m, and trading is closed for the rest of the trading day if a level 3 decline occurred, regardless of the time the CB was triggerd.[251]

The SSCB was replaced by the "Limit up-Limit down" (LULD) pilot plan, in force from April 8, 2013; the first phase rollout of the program included S&P 500 stocks, Russel 1000 stocks and some exchange-traded products (ETPs) (Tier 1 securities), while the second phase rollout, started on August 5, 2013, progressively included all the remaining NMS stocks (Tier 2 securities).[252] Under LULD plan, the price band within which stocks' prices can move is continuously computed by the SIP, since it is responsible for consolidating and disseminating informations on NMS stocks (Rule 603(b)); SIP applies to each stock a percentage parameter above and below a reference price. The reference price is the arithmetic mean price of eligible reported transactions for an NMS stock over the previous 5 minutes, except for periods following openings and reopenings.[253] Furthermore, SIP continuously calculates the trade weighted average price over the preceding five minutes, called Pro-forma referencce price; if the Pro-forma reference price deviates from the current reference price by more than 1% and the current reference price is in effect for at least 30 seconds, the pro-forma reference price replaces the reference one.[254] The percentage parameter applied depends on the tier of the security (Tier 1 and Tier 2) and on its price:

- For Tier 1 stocks with prices greater than 3$, the percentage parameter is 5% above and below the reference price, while for Tier 2 securities the band percentage is 10%; during

[251] Press Release, SEC, SEC Approves Proposals to Address Extraordinary Volatility in Individual Stocks and Broader Stock Market (June 1, 2012), https://www.sec.gov/news/press-release/2012-2012-107htm; Investor Alerts and Bulletins, SEC, Investor Bulletin: Measures to Address Market Volatility (July 1, 2012), https://www.sec.gov/oiea/investor-alerts-bulletins/investor-alerts-circuitbreakersbulletinhtm.html.
[252] Id. 251.
[253]Id. The reference price at the opening of a trading day is the opening price on the Primary Listing Exchange; if this price does not occur in five minutes from the opening (i.e 9:30 a.m.-9:35 a.m.), the reference price becomes the arithmetic mean of eligible reported transactions over the preceding five minutes. The reference price after a trading halt is the reopening price on the Primary Listing Exchange; if this price does not occur in 10 minutes after the reopening of trading, the reference price becomes the last reference price before the trading pause.
[254] Id. 251.

the open and close of the trading day (i.e., 9:30 a.m.- 9:45 a.m. and 3:35 p.m.-4:00 p.m.) the percentage parameter is doubled (respectively 10% and 20%);

- For Tier 1 and Tier 2 stocks with prices between $0.75 and $3, the percentage parameter is equal to 20% of the reference price, except during the open and close, when it is doubled (40%);
- For Tier 1 and Tier 2 stocks with prices less than $0.75, the percentage parameter is equal to the lesser of $0.15 or 75%, except for the open and close, when it is doubled (lesser of $0.30 or 150%).[255]

Differently from the SSCB, the LULD considers quoted prices, not traded prices; thus, the LULD plan halts trade executions on a stock whose quotes are outside the acceptable band, whithout incurring the risk that trades are executed at erroneous prices, such as stub quotes. Furthermore, the LULD plan introduced the Limit State to allow market participants to quickly correct market prices and resume normal trading without triggering a trading pause. A security enters a Limit State when its NBO or NBB is resting on a price band but does not cross it; during Limit State the SIP does not display any reference price or price band and flags those quotes as "Limit State Quotations" and if within 15 seconds the quote is not reverted to a price within the price band, cancelled or executed, the trading in that stock is halted for 5 minutes.[256] A security enters a "Straddle State" when either the NBO is above the limit upper bound or the NBB is below the limit lower bound or both the quotes are outside the price band; SIP flags these quotes as "Non Executable" and stop trading on the relevant side of the market. Differently from the Limit State, the Straddle one has no limit of time: it ends when the quote returns within the price band, when it enters a Limit State, when trading closes or when the Primary Listing Exchange declares a trading pause.[257]
Figure 31 shows a simulation of LULD plan functioning for a particular stock; at the beginning of the trading day the reference price (the red line) is equal to the opening price of $100 and the price bands (the black lines) are equal to 10% above and below the reference price over the first 10 minutes of trading. We can see that the reference price updates at 9:35 a.m., when the pro-forma reference price is equal to $101.10, deviating by 1% from the reference price; after 9:45 a.m. the price band narrows to 5% of the reference price. The stock enters a Straddle State at 10:13:30 a.m., when the NBO (blue line) increases above the upper limit bound and the NBB (yellow line) is within the price band. The straddle state ends at 10:14 a.m., when the stock enters a Limit State (the pink band): the NBO is still above the upper limit bound and the NBB

[255] Id.
[256] Id.
[257] Id.

equals the upper price band. Eventually, trading resumes at 10:19:15 a.m, since both the NBO and NBB are within the price band.[258]

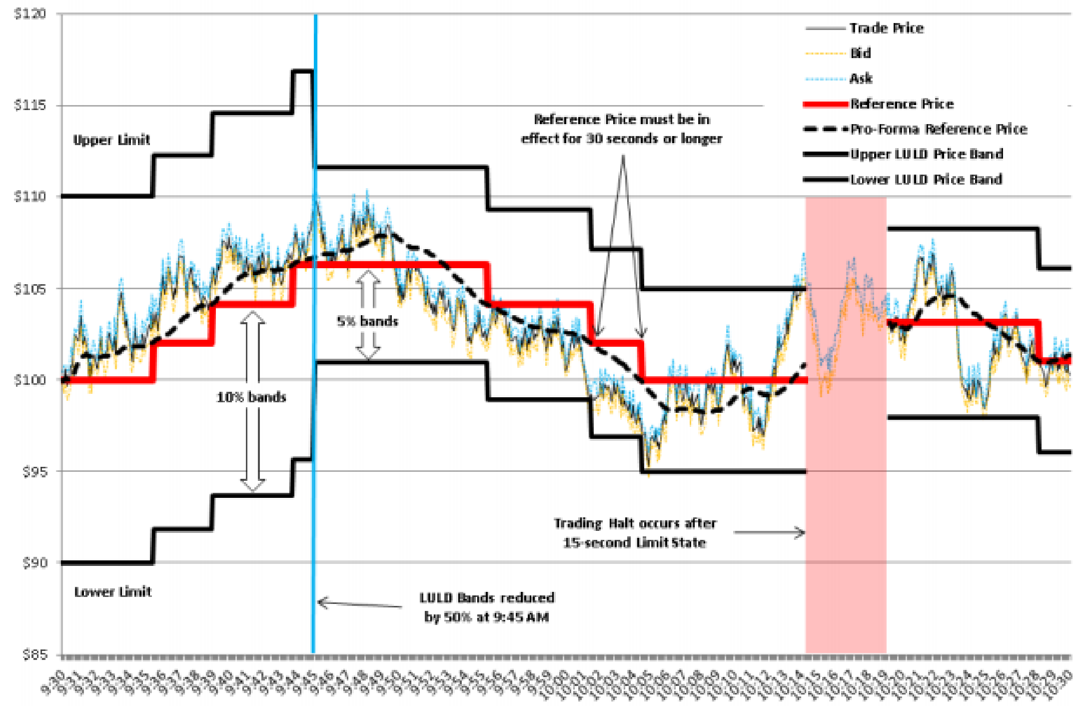


Figure 31: LULD Reference Price and Price Bands Example

Source: Claudia E. Moise, Paca Flaherty (2017)

In Europe, MIFID II requires trading venues to have in place appropriate mechanisms to manage excessive market volatility: the Directive requires trading venues "to have in place effective systems, procedures and arrangements to reject orders that exceed pre-determined volume and price thresholds or are clearly erroneous" and "to temporarily halt or constrain trading if there is a significant price movement in a financial instrument on that market or a related market during a short period".[259] Differently from US regulation, under EU regulation there are no market wide circuit breakers and there is strong heterogeneity in the safeguard mechanisms adopted by EU trading venues. In practice, EU trading venues use two types of volatility safeguards: CB and price collar; the latter does not halt trading as when a CB threshold is crossed, but constraints it: orders that are executed at prices above or below a certain threshold

[258] Id.

[259] Articles 48(4) and 48(5) Directive 2014/65/EU.

or whose sizes exceed a certain limit are rejected while the trading is not halted.[260] Figure 32 shows a typical CB pattern in EU trading venues; when an incoming order is executed at a price outside the price band, it is rejected or partially executed (in case of a large order executed at progressively inferior prices) and the CB is triggered. Once the CB is triggered the trading venue either immediately stop the continuous trading on that stock and open an auction trading session to correct liquidity imbalance or it freezes the order book (no order can be submitted modified or cancelled) before the auction. During the auction phase investors can modify, cancel or submit orders (i.e. called phase) and after a random time interval the orders with matching prices are executed following a single-price auction (i.e. price determination phase), while the others are rejected; finally, continuous trading resumes at the clearing price from the call auction.[261] The CB thresholds, as the duration of trading halt, are determined taking into account a variety of stock characteristics; ESMA provides a non exaustive list of stock characteristics that trading venues have to consider to correctly calibrate their CBs.[262]

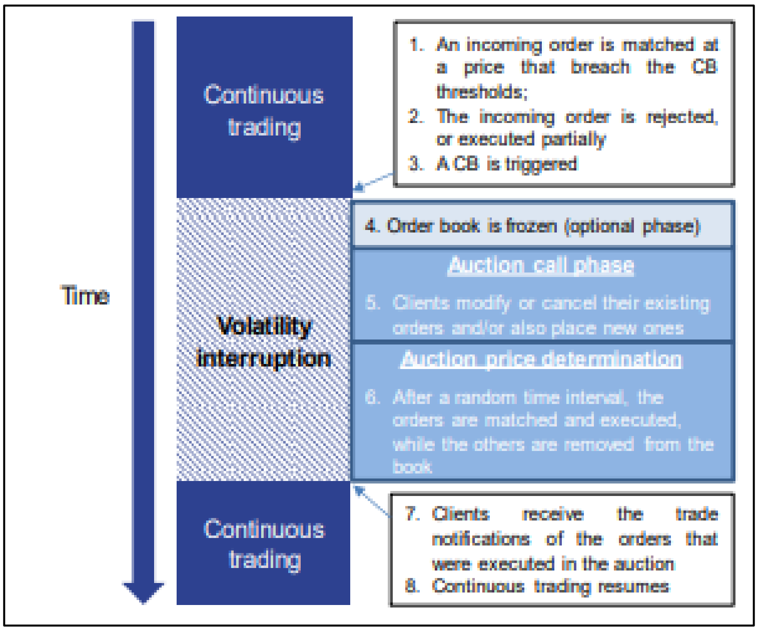


Figure 32: CB mechanism

Source: Cyrille Guillaumie, Giuseppe Loiacono, Christian Winkler, Steffen Kern, (2020)

[260] Cyrille Guillaumie, Giuseppe Loiacono, Christian Winkler, Steffen Kern, 2020, Market impacts of circuit breakers – Evidence from EU trading venues, ESMA Working Paper No. 1.
[261] Id.
[262] ESMA, 2017, Guidelines, Calibration of circuit breakers and publication of trading halts under MiFID II.

Trading venues should calibrate CBs on the liquidity of the financial instrument: more liquid stocks (large cap stocks) should have more stringent price bands than illiquid ones (small cap stocks), since the orders' market impact is larger for the latter. They should calibrate them with respect to expected future events that can affect the liquidity pattern of the stock, such as new issuances or expected corporate actions (for example mergers or acquisitions). The calibration should be supported by a statistical analysis of the financial instrument's historical volatility, taking into account metrics such as absolute maximum intraday deviation, overnight volatility to infer future volatility.[263] The fragmentation of markets requires a coordination of CBs between correlated financial instruments: if the trading halt for a stock is triggered only on one trading venue, the disruptive algorithm can either move to other venues where the same stock is listed (cross-markets) or where associated derivatives or ETPs are listed (cross-assets) or migrate to off-exchange venues, transmitting disruption elsewhere, as we have seen in the May 6, 2010 Flash Crash in section 3.4.5 , where the liquidity crisis was transferred from CME to NYSE. CBs thresholds should depend on the time of the trading day: usually, an higher number of CB thresholds are breached in the first 30 minutes of the trading day, when prices are volatile due to the incoming flow of new informations that they have to reflect, or between 2:30 p.m. and 3 p.m. UTC, when European investors react to informations from the opening of US market.[264]

Also the duration of trading halt is heterogeneous among EU trading venues. ESMA studied CBs' effect on EU trading venues from April 1, 2016 to December 31, 2016 with a sample of 3360 financial instruments (stocks, futures, ETFs, Depositary receipts, foreign exchange rate derivatives); during this period Borsa Italiana and Euronext Brussels had the highest average CB duration, respectively 10 and 12 minutes. Borsa Italiana had also the highest CB's variability, from a minimum of 5 to a maximum of 50 minutes; the dispersion was low for all the other EU trading venues, with a CB average duraton of 4 minutes (Figure 33).[265]

[263] Id.
[264] Id. 261.
[265] Id.

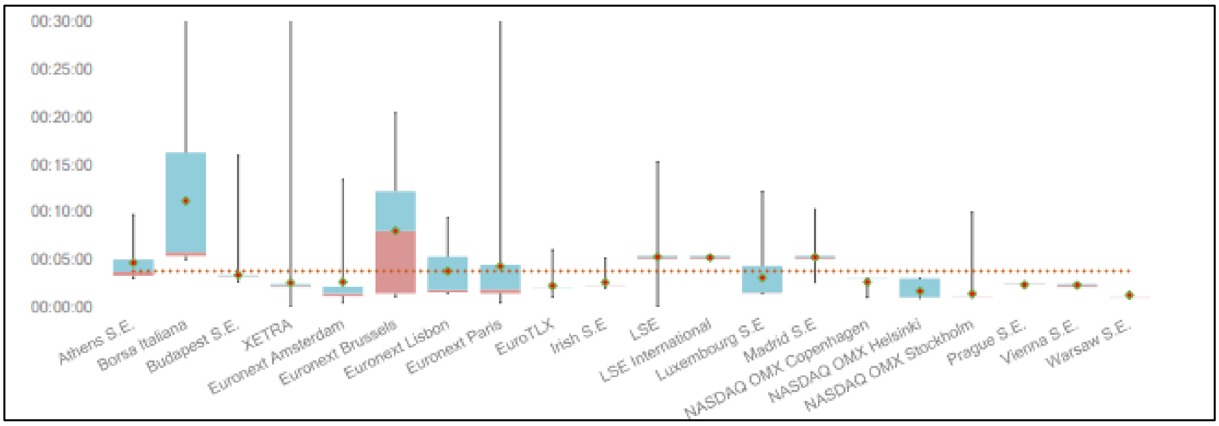


Figure 33: CB heterogeneity across EU trading venues
Source: Cyrille Guillaumie, Giuseppe Loiacono, Christian Winkler, Steffen Kern, (2020)

4.4 Market making obligations

One of the main problem that we have seen studying the effect of HFT on market quality is the lack of dependability of the liquidity provided by HFT market makers; the quality of their liquidity is not the same as that provided by exchange-registered market makers because they are informal market makers, they are not subject to the regulatory obligation of continuously quote two-sided market. Thus, the liquidity that appears in the order book may not be the actual liquidity if HFTs can suddenly withdraw their liquidity, with negative consequences on market participants' transaction costs, expecially for large orders, and market volatility. Financial markets regulators should adress this problem by imposing also to HFT market makers the same affirmative obligations of designated market makers; these market making obligations can include a minimum resting time, size (i.e. 200, 500 or 1000 shares) and market depth (i.e. 3-5 levels below the applicable price obligation) for quotes, quoting securities of a minimum market capitalization (HFTs usually trade large cap stocks) and, as we have seen in the previous section, quoting prices that are within a specific range around the NBBO in order to avoid trades executed at irrational prices (stub quotes)[266]. Efforts towards this direction have been made by EU regulators; MIFID II introduced specific provisions for algorithmic traders and HFTs that engage in market making activity, requiring an investment firm that engages in algorithmic trading and acts as a market maker to "carry out this market making continuously during a specified proportion of the trading venue's trading hours, except under exceptional circumstances, with the result of providing liquidity on a regular and predictable basis to the trading venue".[267] Furthermore, Commission Delegated Regulation EU 2017/578 of 13 June

[266] Id. 226.
[267] Article 17(3)(a) Directive 2014/65/EU.

2016 (RTS 8) supplements MIFID II specifying the requirements of market making agreements and schemes; it requires a market makers to post quotes, with respect to the financial instrument or instruments in which they pursue a market making strategy, of comparable sizes and competitive prices and that these quotes remain in the order book for at least 50% of the trading day, excluding opening and closing auctions.[268] In particular, the provision specifies that the term comparable size means the quotes on the two sides of the book do not diverge by more than 50% from each other, while the term competitive prices means that quotes' prices are posted at or within the current NBBO range.[269] Competitive and comparable quotes and quote minimum resting time respectively help reducing market volatility and improves liquidity provision by ensuring that quotes will remain available for minimum time, also during market stress.

MIFID II specifies market makers can withdraw posted liquidity under exceptional circustances, that are laid down by ESMA in RTS 8:

- Extreme market volatility that triggers CBs on financial instruments market makers continuously quote or on its derivatives;
- War, industrial action, civil unrest or cyber sabotage;
- Disorderly trading conditions caused by the incapacity of the trading venue to maintain fair, orderly and transparent execution of trades: the performance of the trading venue system is affected by delays and interruptions, multiple erroneous orders or transactions and the incapacity of the trading venue to provide services;
- Investment firm cannot conduct prudent risk management practice because of technologicl issues (problem of data feeds or other systems) essential for its market making strategy or risk management issue (regulatory capital, margining and access to clearing).[270]

The aforementioned market making obligations can have the opposite effect to cause the HFTs to exit the market if the cost to pursue marketing making strategies is not adequately rewarded. As we have seen in section 2.1.2, trading venues compete for order flows adopting maker-taker fee structure; through this pricing model, trading venues pay liquidity providers a rebate for posting liquidity while charge a fee to traders that take liquidity. From the debate on the maker-taker pricing model arisen the proposal to make this model dynamic; for instance, Blackrock, one of the larger asset manager of the world, proposed to calibrate the pricing model to the

[268] Article 1(1)(a)(b) Delegated Regulation 2017/578/EU.
[269] Article 1(2)(c)(d). Delegated Regulation 2017/578/EU.
[270] Article 3 Delegated Regulation 2017/578/EU.

capitalization of the stock: clearly, illiquid assets as small cap stocks need more liquidity than large cap stock, thus trading venues should offer higher rebates to market makers providing liquidity for this class of stocks.[271] Also BATS proposed to SEC to promote a dynamc maker-taker model, with rebates calibrated to stocks' characteristics such as market capitalization, average daily volume, inclusion in certain stock indexes. In particular, BATS proposed to implement tiered market access fees, reducing the access fee cap, and its associated rebate, to $0.0005 per share (five cents per 100 shares) for highly liquid stocks, from the current $0.003 (17 C.F.R. 242.610); on the contrary, moderately liquid and illiquid stocks should have access fee and associated rebates from $0.0005 per share.[272]

Following the same reasoning, dynamic maker-taker model should be implemented to reward HFT market makers to post liquidity also during highly volatile period: to incentivize HFTs to continuously quote two-sided market during period when the inventory risk significantly increases, as during excess market volatility, trading venues should pay them an higher rebate when trading activity triggers a circuit breaker or exceeds a limit-up limit-down band.

4.5 Unfiltered/naked access prohibition

In section 1.3.2 we have studied sponsored access as a direct electronic access that allows market participants to directly trade on the trading venue without the intermediation of a member and the use of its infrastructures, but only using member's market participant identification (MPID). In particular, the unfiltered, or "naked", access enables traders to enter the marketplace without any pre-trade risk controls. Clearly, this form of direct electronic access drastically reduces trading latency and even if it can be beneficial for all kind of traders, it is of particular importance for HFTs, since they have the technological advantage to fully leverage the profit opportunities offered by this low-latency infrastructure. However, the use of sophisticated and high-frequency technology combined with the capacity of traders to place orders without the intermediation of a broker has risen concerns on the quality of the risk controls put in place by broker-dealers offering this service. The absence of appropriate pre-trade risk controls can increase the risk that customers will enter erroneous orders as a consequence of computer glitches or human errors, fail to comply with specific regulatory requirements and breach a credit or capital limit.[273] The detrimental consequences on financial

---

[271] U.S. Equity Market Structure: An Investor Perspective, VIEWPOINT (BlackRock, New York, N.Y.). Apr. 2014.

[272] Open Letter, BATS, Market Structure Reform Discussion, at 3-4 (Jan. 6, 2015).

[273] SEC, SEC Adopts New Rule Preventing Unfiltered Market Access (Nov. 3, 2010), available at https://www.sec.gov/news/press/2010/2010-210.htm.

stability increases if the unfiltered access is provided to HFTs, since the high-frequency algorithms they use can seriously destabilize the market in the occurrence of a glitch or of an high-frequency trading strategy not adequately controlled by capital limit; furthermore the risks of naked access has become more severe with the high interconnection of financial markets and with the correlation of trading strategies adopted by algorithmic trading, which can exacerbate the systemic risk.

On the wave of these concerns, on November 2010 the SEC prohibited unfiltered access; the adopted Rule requires broker-dealer that provides market access to an exchange or ATS through the use of its MPID to a customer to "establish, document, and maintain a system of risk management controls and supervisory procedures reasonably designed to manage the financial, regulatory, and other risks of this business activity".[274] The financial risk management controls and supervisory procedures must be designed by the broker-dealer that provides the market access to prevent the entry of orders that exceed pre-set credit or capital thresholds and the entry of erroneous orders, rejecting those that exceed pre-determined price or size parameters.[275]

MIFID II follows the same pattern of US regulation with respect to unfiltered sponsored access; the Directive requires investment firms providing direct electronic access to have in place effective systems and controls which ensure "a proper assessment and review of the suitability of clients using the service, that clients using the service are prevented from exceeding appropriate pre-set trading and credit thresholds, that trading by clients using the service is properly monitored and that appropriate risk controls prevent trading that may create risks to the investment firm itself or that could create or contribute to a disorderly market or that could be contrary to the market abuse regulation (section 4.8)".[276]

4.6 Taxation of HFT activity

HFT manipulative strategies such as quote stuffing or spoofing are based on the capacity of HFTs to quickly cancel their fleet quotes before non-HFTs can fill them; given the detrimental effect of these trading strategies for non-HFTs' transaction costs, one solution to this problem should be a taxation of HFT activity: as any other tax, financial transaction taxes reduce the amount of the taxed activity.[277] The taxation of the trading activity on the base of the messages activity of market participants should disincentivize HFTs to flood the market with thousands of orders, forcing them to trade more on economic fundamentals than on the speed of the ultra-

[274] 17 CFR § 240.15c3-5(b).
[275] 17 CFR § 240.15c3-5(c)(1).
[276] Article 17(5) Directive 2014/65/EU.
[277] Id. 226.

fast data feeds they use. However, we have seen in section 1.5.1 that HFT activity identification is not a straightforward task; in particular, the direct approach can lead to an underestimation of the HFT activity, which would result in a lower tax base and on a lower effect on HFT activity, while the indirect method which leads to an overestimation of HFT activity, encompassing in its scope also non-HFT activity, which increases the tax base but at the same time jeopardize the validity of the tax.

Italy was the first country to adopt a taxation on HFT activity, the Italian Financial Transaction Tax (IFTT), adopted in february 2013.[278] The provision identifies HFT activity any trading activity that fulfills the following features:

- It is generated by a computer algorithm that automatically determines decisions related to the sending, modification (price and size) and cancellation of orders and relevant parameters;
- The time between placing a buy or sell order and subsequent modification or cancellation of the same order by the same algorithm does not exceed half a second.[279]

The tax rate is equal to 0.2% and it is daily computed when the ratio between the sum of daily cancelled and modified orders and the sum of daily submitted and modified orders is over 60%, only considering orders cancelled and modified within half a second[280]. Since the provision has been introduced to limit the adverse effect of HFT on market integrity and stability, it does not affect HFT market makers, whose positive activity is not considered in the computation of the tax base.[281]

Althought the tax was targeted for HFT aggressive strategies, a negative consequence of this tax should be a drastic reduction of HFT activity, which leads to a reduction of market liquidity: in order to avoid the tax, HFTs will allocate their operations to other exchanges under different jurisdictions; this liquidiy reduction can be particularly detrimental for the non-well developed italian capital market. The year Italy’s financial transaction tax was introduced, trading in italian stocks fell by 34,2% and the government raised €200 millions against an expected revenue of €1 billion.[282] Furthermore, studies have found that volatility and bid-ask spread significantly increased after the tax introduction.[283]

---

[278] Article 12 of the Law Decree of the Italian Minister of Economy and Finance of 21 February 2013.
[279] Id.
[280] Article 13(1) of the Law Decree of the Italian Minister of Economy and Finance of 21 February 2013
[281] Article 12(1)(a)(1) of the Law Decree of the Italian Minister of Economy and Finance of 21 February 2013.
[282]Maria Coelho, 2016, Dodging Robin Hood: Responses to France and Italy’s Financial Transaction Taxes, University of California, Berkeley, College of Letters & Science, Department of Economics
[283] Tobias Ruhl & Michael Stein, 2014, The Impact of Financial Transaction Taxes: Evidence from Italy, 34 Economic Bulletin 25, 32.

Thus, the success of a tax on financial transactions depends on an international regulatory coordination between countries, otherwise, as we have seen for trading venues fragmentations, HFT firms will pursue arbitrage strategies leveraging regulatory fragmentation among jurisdictions.[284]

4.7 Equal Market Access

4.7.1 Trading venue's speed bump

HFTs' structural insider dealing has brought back into capital markets the uneven playing field financial regulators have tried to limit with provisions prohibiting insider dealing (section 3.7); thus, exchanges have long debated on how to restore investor confidence directly intervening to restrict the speed advantage of HFTs in order to equalize the access to public informations to all market participants. Since latency arbitrage strategies are based on the use of exchanges' direct feeds and co-location servers to first receive market data and algorithms that automatically transact on these informations, the solution to the two-tiered market seems to pass through restricitons on the use of one of these device, for example reducing the informations conveyed by exchange's direct feeds or prohibiting to co-locate traders' own servers near exchanges' matching engine. HFTs should suffer some losses since direct feeds are less in-depth and not sufficiently informative to give them a consistent edge on market order flows' dynamics.[285] However, these restrictions can hurt also non-HFTs, since the use of direct feeds is not limited to HFTs but include a larger range of traders, which encompasses also traders that not necessariliy use this infrastructure to conduct trades at ultra-fast speed. Furthermore, prohibiting co-location services would not refrain HFTs to buy real estate as close as possibe to an echange.[286] The restore of an equal access to the market needs a larger structural approach; in the wave of these considerations, some exchanges' operators started to adopt speed bumps to slow down the message traffic on its trading venues. The first trading venue which adopted this mechanism was Investor Exchange (IEX); born as a dark pool alternative trading system in October 2013, on April 2015 it began to operate as a national securities exchange and on June 2016 it introduced a 350-microsecond speed bump to delay all incoming and outgoing messages to and from market participants.[287] The 350 microseconds delay of messages, which includes market orders, limit orders, order cancellations, trade executions and quote messages reported on exchange's direct feeds, but not on the SIP, mitigates the speed advantages of HFTs.

---

[284] Id. 226.
[285] Id. 215.
[286] Id.
[287] Edwin Hu, 2018, Intentional Access Delays, Market Quality, and Price Discovery: Evidence from IEX Becoming an Exchange, SEC Division of Economic and Risk Analysis' Working Papers.

For instance, the structural delay reduces HFTs' front-running profits: when an HFT will send ping orders to detect the presence of a large order in order to trade ahead of it on other exchanges towards which it has sent the other child orders, the HFT will not be aware of the order executions before 350 microseconds, allowing the other orders to be executed on other exchanges without being front-run.

The approval of the IEX's 350 microseconds delay required SEC to give a new interpretation of "protected quotations" of Order Protection Rule; the provision establishes that trading centers must prevent trade-through of protected quotations; Rule 600(b)(62) defines "protected quotations" as a "protected bid or a protected offer".[288] Rule 600(b)(61) in turn defines a "protected bid and a protected offer" as a quotation in an NMS stock that is an "automated quotations" that represents the best bid and the best offer of a national securities exchange or a national securities association.[289] In particular, Rule 600(b)(37) defines automated quotation as one that permits an incoming order to be immediately and automatically executed; thus, the 350 microseconds delay of IEX was in contrast to this definition.[290] However, the SEC allowed IEX structural delay; in June 2016, the SEC issued a new interpretation of Rule 611 to allow for "de minimis" intentional access delay to automated exchange quotations: the Commission stated that a delay of less than one millisecond (1000 microseconds) in quotation response would not impair a market participant's ability to access a quote.[291]

The new SEC interpretation of Rule 611 led to IEX copycats; in October 2017 the SEC approved a plan by Chicago Stock Exchange similar to that of IEX; the CHX introduced a 350 microseconds speed bump as IEX, but while IEX slowed down every market participants trading on its venue, CHX slowed down only liquidity demanders, those who trade against standing limit orders posted by market makers.[292] Non marketable limit orders and cancel messages for resting orders were immediately processed without delay; this asymmetric speed bump gives market makers a 350 microseconds window to modify resting limit orders before executing it, thus reducucing the probability an HFT pick off their orders at stale prices. Thus, the CHX "Liquidity Taking Access Delay" discourages latency arbitrage and incentivizes market makers to make tighter and deeper markets; however, as seen for HFT ghost liquidity,

---

[288] 17 CFR 242.600(b)(62).
[289] 17 CFR 242.600(b)(61).
[290] 17 CFR 242.600(b)(4).
[291] SEC, Commission Interpretation Regarding Automated Quotations Under Regulation NMS (2016), available at https://www.sec.gov/rules/interp/2016/34-78102.pdf.
[292] Matt Levine, Speed Bumps Are the Hot New Thing for Exchanges, August 31, 2016, https://www.bloomberg.com/opinion/articles/2016-08-31/speed-bumps-are-the-hot-new-thing-for-exchanges; Alexander Osipovich, SEC Approves Chicago Stock Exchange's 'Speed Bump' for Trading, October 19, 2017, https://www.wsj.com/articles/sec-approves-chicago-stock-exchanges-speed-bump-for-trading-1508453511.

the liquidity that appears on the order book is illusory, since the market makers can cancel or modify their resting orders if the market moves against them.[293]

4.7.2 Batch Auction

A recent study conducted by professors Eric Budish, Peter Cramton, and John Shim (2015) has highlighted how the HFT arms race is a sympton of a basic flaw in the design of modern financial markets: the continuous-time trading; that is, the continuous limit order book allows market participants to buy or sell stocks or other financial instruments at any instant during the trading day, and in particular at ever smaller time intervals.[294] The main finding of Budish et al. was the correlation breakdown at the high-frequency time horizons under continuous-time trading; they studied the correlation between the two most traded financial instruments tracking the S&P 500 index, the SPDR S&P 500 ETF (SPY) and the S&P 500 E-mini futures contract (ES) over a sample period from 2005 to 2011. As we have seen studying the May 6, 2010 Flash Crash, since these two financial instruments track the same underlying, their correlation should be high and constant given the near-arbitrage relationship between them; panels (a) and (b) of figure 34 display the median (solid line), minimum and maximum (dotted lines) correlation between the return of the ES and SPY bid-ask midpoints as a function of the return time interval in 2011, with time intervals from 1 millisecond to 60.000 milliseconds (60 seconds) over all trading days in 2011.[295]

[293] Id.
[294] Eric Budish, Peter Cramton, and John Shim, 2015, The high frequency trading arms race: frequent batch auctions as a market design response, Quaterly Journal of Economics, Vol.130 (4), 1547-1621.
[295] Id.

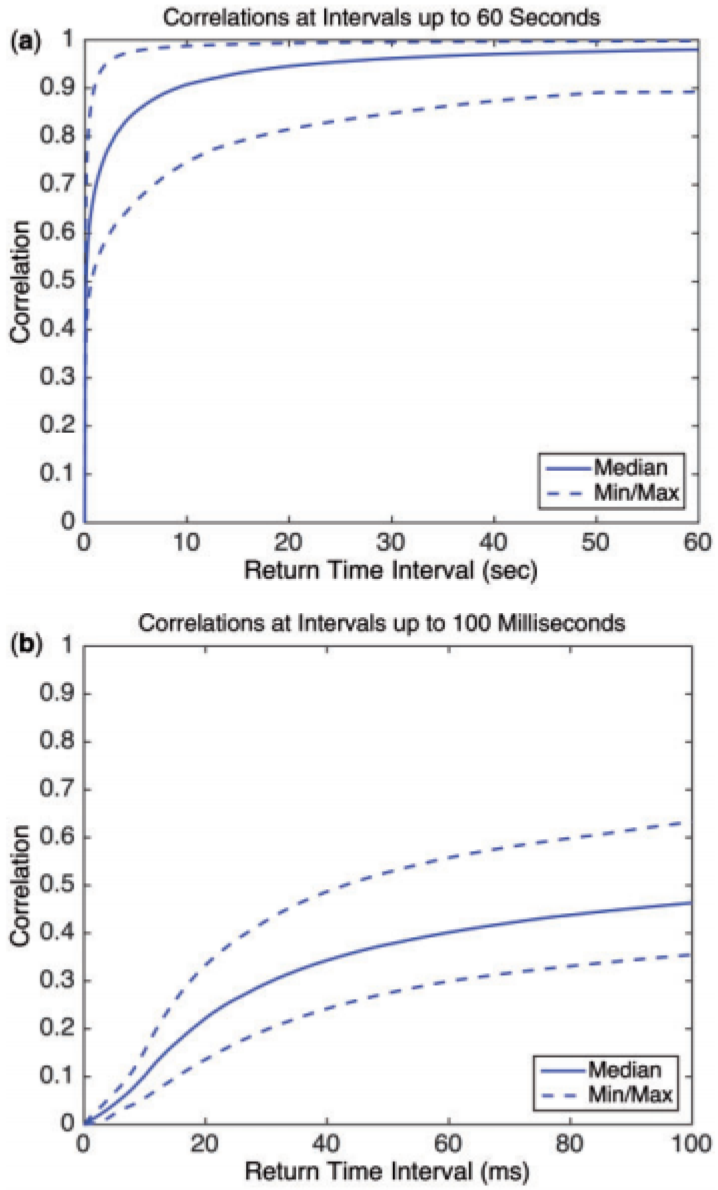


Figure 34: ES-SPY return correlation
Source: E. Budish et. al (2015)

From the figure we can see that the correlation between ES and SPY return increases with the return time interval, approximating to 1 for a long enough interval (at a 60 second return time interval the correlations is almost 1); on the contrary, the correlation breaks down at high-frequency time horizons: the 10 millisecond correlation is just 0,1016, while the correlation at 1 millisecond time interval drops to 0,0080.

The correlation breakdown at high-frequency horizons gives arbitrage opportunities between the two instruments for those HFTs who are fast enough to exploit them. However, Budish et al. found that the HFT arms race did not affect the size of the arbitrage premium but it just continuously raised the bar of speed HFTs have to reach in order to capture it; during the sample period 2005-2011, the duration of arbitrage opportunities significantly declined, from a median of 97 milliseconds in 2005 to a median of 7 milliseconds in 2011, while the profitability of ES-SPY arbitrage opportunities remain almost constant during the sample period, with a median profit at around 0.08 index point per contract traded.[296]

The authors proposed a new market model to solve the intrinsic correlation breakdown of continuous-trading market: a discrete-time trading market in which trading day is divided into extremely frequent discrete time intervals, for example 100 milliseconds; the new market model is called “Frequent batch auction” market. Instead of continuously matching all the incoming market orders with the resting ones as they arrive, the exchange collects all the orders submitted within a specic time interval and then executes them through a single-price auction; the orders that are not executed (their bids and offers are respectively lower and higher than the clearing price) are posted in the following batch auction with the new orders that will be submitted.[297]

The new market design should mitigate the HFT arms race, shifting the competition from speed to price, since low-latency trading no longer ensure trade execution in a batch-auction market. Thus, instead of heavily invest in ever faster infrastructures, batch auction markets should force HFTs to spend more time and resources in securities’ fundamental researches, dissuading them from conducting aggressive strategies that add little to price discovery (ghost liquidity strategies).[298]

As for the debate over the adoption by the trading venues of structural delay mechanisms, also the application of a discrete-time trading market seems to be in contrast with the current US regulation. The most importan principle of capital markets regulation is the “best execution” of investors’ order, which require broker-dealers to route the clients’ order to the trading venue quoting the NBBO (Rule 611). In a continuous-trading market broker-dealers are sure to comply with the order protection rule, since their orders are immediately executed at the NBBO when they arrive to the exchange that is posting it; however, in a batch-auction market the execution price is determined at the end of a batch auction, thus broker-dealers cannot know in advance which trading venue will post the NBBO.[299]

The Chicago Stock Exchange was the first trading venue that attempted to incorporate batch auctions into US equity markets; in October 2015 the SEC approved the Chicago Stock Exchange’s plan to launch an on-demand batch auction platform called CHX SNAP (Sub-

[296] Id.
[297] Id.
[298] Id. 226.
[299] Id.

second Non-displayed Auction Process) to promote block trading.[300] Differently from the market model proposed by Budish et al., CHX SNAP is an on-demand batch auction market, which means that batch auctions are activated by market participants. CHX introduced the SNAP platform to protect large investors from being front-run by HFTs; it introduces a new type of order, the Start SNAP order, which initiates a SNAP auction cycle if submitted.[301] Once the SNAP auction cycle is activated, the exchange suspends automatic execution of orders in the relevant security, notifies to all market participants that a SNAP cycle begun and suspends dissemination of any information about the relevant security; in particular, as a dark pool, it does not reveal the price and size of the order, depriving HFTs of the trades' information leakages they use to trade ahead institutional investors. When the SNAP cycle begins, the SNAP CHX book is established during a period ranging from 475 to 525 microseconds during which traders can participate to the batch auction submitting SNAP eligible orders.[302]Furthermore, during the SNAP cycle, no submitted orders can be cancelled, thus traders are incentivized to send only bona fide orders. At the end of this stage, the matching system determines the SNAP price, that is the single-clearing price that maximizes the trading volume in the subject security; if the minimum order size requirement is not met, the SNAP cycle is closed and the exchange comes back to continuous trading. On the contrary, if the size requirement is met the SNAP eligibe orders are executed at the SNAP price; the more aggressive orders on the CHX SNAP book are routed to exchanges posting protected quotations that improve the SNAP price, in accordance with the trade-through prohibition under SEC Regulation; CHX SNAP also routes SNAP orders to other exchange posting the same SNAP price in order to fill order imbalances within the CHX SNAP book at the SNAP price. At the conclusion of the order matching stage, continuous trading on the subject security is restored and all the wholly or partially unexecuted orders migrate from SNAP book to CHX book. [303]

[300] Order Approving CHX Proposed Rule Change to Implement Intra-Day and On-Demand Auction Service, Exchange Act Release No. 76087, 80 Fed. Reg. 61540, (Oct. 6, 2015), https://www.gpo.gov/fdsys/pkg/FR-2015-10-13/pdf/2015-25886.pdf.

[301] Id. In order to be eligible as a start SNAP order the order must fulfill some requirements with respect to size and price: the size must be at least 2500 shares with a minimum aggregate notional value of $250.000 or at least 20.000 shares with no minimum notional value; the limit price must be at or through the NBBO.

[302] Rebecca Lewis, 2015, A new approach to stock market execution, The Federal Reserve Bank of Chicago, Chicago Fed Letter, Number 343. Given the very short period over which traders can submit SNAP orders, non-HFTs can submit a specific order type, the so called SNAP auction only order (AOO) at every time during the trading session: this order will be submitted to the CHX SNAP book when a SNAP cycle begin.

[303] Id.

4.8 HFT Market manipulation regulation

In section 3.6 we have outlined the importance of transparency for capital market efficiency: listed companies provide the market with a rich book detailing their inner workings, informing investors with routine updates about their activities, organizational changes and economic performance. The mandatory disclosure of listed companies are one of the main tool to ensure market integrity, essential for investors' trust on how the assets are valued. However, corporate disclosure is not sufficient to safeguard market integrity but financial regulators have to continuously monitor capital markets in order to detect trading activity intentionally conducted to give a misrepresentation of securties' prices, better known as market manipulation activities. The Regulation (EU) 596/2014 of the European Parliament and of the Council of 16 April 2014, well known as Market Abuse Regulation, the regulation on market abuse practices on EU financial markets, defines market manipulation as a trading acticity that "gives, or is likely to give, false or misleading signals as to the supply of, demand for, or price of, a financial instrument, a related spot commodity contract or an auctioned product based on emission allowances"[304]. Although market manipulation practices have not began with the rise of HFT, the regulation points out that the definition of market manipulation has to be adapted to "new forms of trading or new strategies that may be abusive", further specifying that, in the face of an increasingly automation in the trading of financial instruments, it is necessary to define market manipulation also providing "examples of specific abusive strategies that may be carried out by any available means of trading including algorithmic and high-frequency trading".[305]

In accordance with the technological innovations of securities' trading, the regulation specifically adresses HFT expanding the definition of market manipulation in order to encompass strategies where the placement of a large number of orders is intended to manipulate prices, rather than any actual trading; thus, it provides that it is considered market manipulation trading activities that consist in "the placing of orders to a trading venue, including any cancellation or modification thereof, by any available means of trading, including by electronic means, such as algorithmic and high-frequency trading strategies" that have the effect of:

- disrupting or delaying the functioning of the trading system of the trading venue or being likely to do so;

[304] Article 12(1)(a)(i) Regulation 596/2014.
[305] Recital 38 Regulation 596/2014.

- making it more difficult for other persons to identify genuine orders on the trading system of the trading venue or being likely to do so, including by entering orders which result in the overloading or destabilisation of the order book; or
- creating or being likely to create a false or misleading signal about the supply of, or demand for, or price of, a financial instrument, in particular by entering orders to initiate or exacerbate a trend.[306]

The list of abusive behaviors provided by the regulation has the effect of prohibiting HFT strategies that fall within the scope of the provision, such as quote stuffing, pinging, pump and dump and spoofing. All these strategies existed before the advent of HFT; the problem is that HFT can increase the amount of manipulation practices implementing them on a larger scale and over a shorter period than ever before, which allow them to exit the market before other traders are able to react and make more difficult for regulators to detect their manipulative trading.[307]For instance, we know that HFT is characterized by an high order-to-trade ratio since they cancel most of their posted orders at very high frequency; this cancellations are due to their continuous update of market data, in particular when they engage in market making activity; however, this practice can be also effective in pinging or layering the market by simulating liquidity dynamics. Thus, the use of consolidated audit trail such as MIDAS or CAT are essential for reconstruct the overwhelming activity of HFT in order to detect manipulative strategies.

4.8.1 Spoofing and layering

We have already introduced the aforementioned manipulative trading strategies in section 2.5, under the category of “ghost liquidity strategies”. The most used manipulative strategy carried out by HFTs is spoofing, that is specifically adressed by US regulation. In 2010, Section 747 of the Dodd-Frank Wall Street Reform and Consumer Protection Act (the “Dodd-Frank Act”) amended section 4c(a) of the Commodity Exchange Act (CEA) to specifically target HFT. Section 4c(a) of the CEA (7 U.S.C. 6c(a)) identifies prohibited transactions, that are those that are recognized as “wash sale” or “accomodation trade”, that represent a fictious sale or that are used to cause any price to be reported, registered or recorded that is not a true and bona fide price.[308] Section 747 of the Dodd-Frank Act amendment adds to this section new provisions on

[306] Art. 12(2)(c) Regulation 596/2014.
[307] Regulatory Issues Raised by the Impact of Technological Changes on Market Integrity and Efficiency, IOSCO (2011).
[308] 7 U.S. Code § 6c(a)(2).

disruptive practices, prohibiting "any person to engage in any trading, practice, or conduct on or subject to the rules of a registered entity that:

- Violates bids or offers;
- Demonstrates intentional or reckless disregard for the orderly execution of transactions during the closing period; or
- Is, is of the character of, or is commonly known to the trade as "spoofing" (bidding or offering with the intent to cancel the bid or offer before execution)."[309]

Thus, US regulation specifically bans spoofing. Spoofing is a particular momentum ignition strategy that try to manipulate the price of an asset submitting non-bona fide limit orders on one side of the market, usually inside the current NBBO, in order to artificially create a buying or selling pressure with the aim to narrow the quoted bid-ask spread; when the flickering orders successfully entice other market participants into following the direction of the interest pressure, the spoofer quickly cancel the submitted limit orders before they are executed and, at the same time, submits market orders on the opposite side of the market that will be executed at the post-spoofing better price.[310] Many times regulators use the terms spoofing and layering interchangeably; even if the purpose of the two strategies are identical, layering is a variant of spoofing: while the former is conducted placing a few large visible non-bona fide orders at prices slightly better than the NBBO, layering implies the submission of a larger number of consecutive non-bona fide limit orders at different level to simulate a selling or buying pressure[311]. Figure 34 gives an example of how a spoofing strategy works. Suppose that the current NBO of stock X is equal to $103.50 and that an HFT spoofer wants to buy it at a better price; the spoofer starts to submit large limit orders within the NBBO (with limit sell prices lower than the current NBO) and increase the market depth on the sell side of the order book to simulate a selling pressure. The spoofer's posted limit sell orders at decreasing prices give other traders the impression of a bear market on stock X, istigating them to follow the trend; thus, the HFT quicly cancels their limit sell orders and submits market buy orders that are executed at the deflated stock price of $96. Once the other traders realize that the stock price is overvalued with respect its fundamental, they began to buy it reverting the trend; furthermore, the HFT can accelerate the reverting process of the stock price by replicating the spoofing strategy on the other side

[309] 7 U.S. Code § 6c(a)(5).
[310] News Release, FINRA, FINRA Joins Exchanges and the SEC in Fining Hold Brothers More Than $5.9 Million for Manipulative Trading, Anti-Money Laundering, and Other Violations (Sept. 5, 2012), available at https://www.finra.org/media-center/news-releases/2012/finra-joins-exchanges-and-sec-fining-hold-brothers-more-59-million
[311] Id.

of the market, by placing large limit buy orders. Once the stock price is reverted to its pre-spoofing price, the HFT spoofer closes its position at a profit, while the other slow-traders suffer a loss of equal size.[312]

However, because the provision on spoofing relies on the intentionality of the trader to carry out the prohibited practice, the enforcement of the rule against an HFT can result challenging. To clarify the interpretation of the Dod-Frank Act Section 6c(a)(5)(c), the CFTC has released guidance on the intent requirement, providing that a market participant has to "act with some degree of intent, or scienter, beyond recklessness to engage in the "spoofing" trading practices prohibited by CEA section 4c(a)(5)(c)".[313]

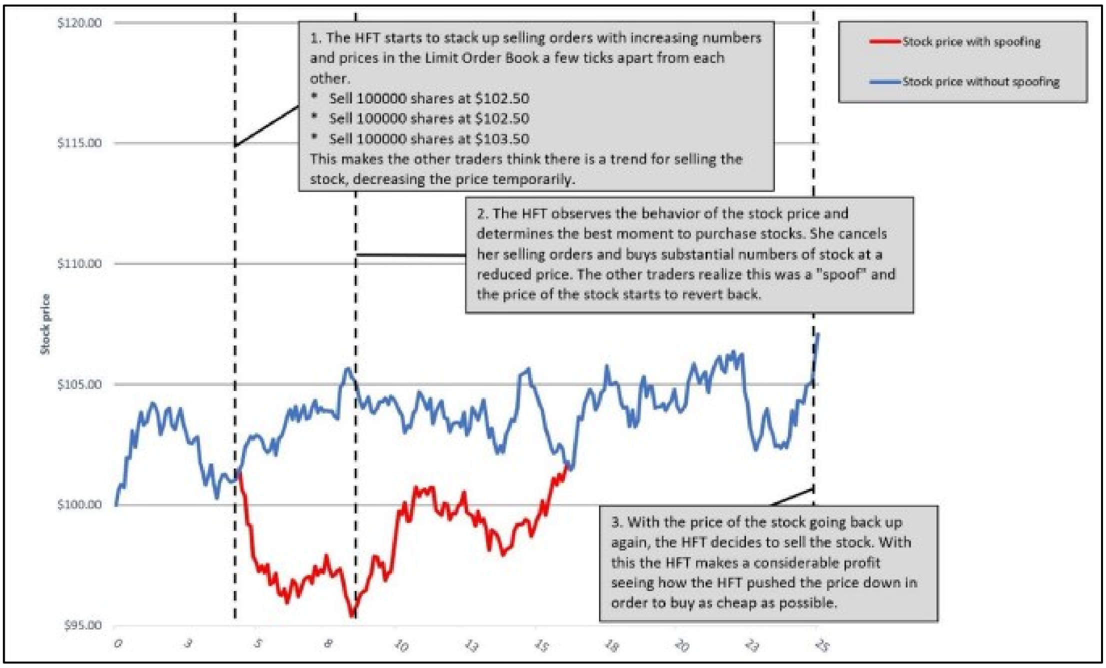


Figure 35: Example of spoofing strategy
Source: Poblocka (2019)

[312] Poblocka, Iwona, 2019, High Frequency Trading: highway to financial hell or economic salvation: a comprehensive review of the High Frequency Trading literature, University of Twente, Industrial Engineering and Management MSc.
[313] U.S. Commodity Futures Trading Commission, Antidisruptive Practices Authority, 78 C.F.R, issue 102 (2013).

CONCLUSIONS

High frequency traders are algorithmic traders that use low latency infrastructures provided by stock exchanges as direct market feeds or co-proximity services to trade at high speed in order to profit from an ever larger number of trading opportunities before others can react to a financial market's information flow that run across traders' screen within time frame in the order of milliseconds. The evolution of algorithmic trading and HFT and their dominance in the trading industry have spurred finance scholars and financial regulators to investigate the consequences of this new breed of traders for market quality and financial stability and integrity.

Many studies have demonstrated the beneficial effects of HFT in terms of market liquidity, volatility and efficiency. The capacity of HFTs of quickly react to incoming informations and the computational power of their computer algorithms to collect and process large amount of market data have made them the best market makers, replacing the traditional exchange-registered market makers, given their capacity to reduce the adverse selection risk from trading with informed traders and thus of quoting a narrower bid-ask spread. For the same reason HFTs are able to reduce market volatility, limiting and reducing excessive price fluctuations filling the order book with liquidity at an higher and ever tighter number of price levels. Furthermore, there is a strong empirical evidence of the HFT's positive effect for financial market efficiency: the capacity of HFTs to analyze the market quicker than non-HFTs allows them to accelerate and improve the price discovery process, also correcting temporary liquidity imbalances following contrarian trading with respect to uninformed trading.

Despite the aforementioned benefits, HFTs' consequences for financial markets change when academics shift their attention from passive trading strategies to aggressive ones; on the light of this concern, one of the main problem arisen with the evolution of HFT is the phenomenon of ghost liquidity, with trading strategies as quote stuffing or layering: the liquidity provided by HFTs are not of the same quality of that supplied by designated market makers: many times HFTs flood the market with "flickering" quotes that they immediately cancel with the aim to not execute them but instead to convey to the other market participants false representation of market dynamics in order to instigate them to trade in a desired direction; consequently, this phenomenon has the effect to increase the transaction costs, since from the zero-sum game of these strategies the profits of HFTs correspond to losses for slow traders.

Detractors of HFT, supported by the harsh criticism against this trading model by Michael Lewis in its famous book "Flash boys", have also pointed out how they have created a two-tiered market, where HFTs can systematically beat slower traders leveraging flash facilities, direct market feeds and co-located exchanges' services, bringing to light the same uneven playing field generated by illegal practices as front running and insider trading.

The events of the May 6, 2010 Flash Crash has shed light on the detrimental effect of HFT for financial stability during period of high volatility. Even if HFTs supply cheaper liquidity, there is a lack of dependability of this liquidity because they are not subject to any affirmative obligation: thus, HFTs tend to withdraw liquidity during period of high volatility, when the provision of liquidity is more needed, contributing to exacerbate the financial instability. This is what happened on May 6, 2010: even if the crash was not generated by HFTs, they contribute to exacerbate the excessive volatility of that day, quickly withdrawing their liquidity and contributing to the first but not the last flash crash, a market crash that differently from those of the pre-HFT era recovered its value within a few minutes.

On the wave of the May 6, 2010 Flash Crash and with the aim to restore investors' protection and confidence, US and EU financial regulators progressively updated securities industry regulation to encompass the HFT phenomenon. They increase trading monitoring and oversight adopting more stringent rule as the US "Large trader reporting rule" and requiring more detailed trading informations from traders engaging in algorithmic trading, limiting the asymmetric information with respect to the algorithmic models implemented.

In order to avoid excessive market volatility as that occurred during the Flash Crash, US and EU regulation introduced or improve trading venues' control systems as circuit breakers or price collars, also requiring HFTs to test their algorithms and to adopt internal risk controls to avoid to generate destabilizing rogue algorithms. Furthermore, financial regulators explicitly introduced provision specifically tailored to prevent market manipulation conducted at high frequency, prohibiting HFT strategies as quote stuffing or spoofing.

Beside the new financial regulation, some trading venues and jurisdictions have adopted rules and structural reforms to limit the activity of HFTs; the IEX and Chicago Stock Exchange introduced a 350 microseconds speed bump to delay the HFT message activity from and to the trading venues in order to reduce the adverse selection risk of slow traders, while Italy was the first country to introduce a financial transaction tax on the message activity of the HFTs. Furthermore, some trading venues have started to experiment a new market design based on a discrete-time trading, the so called batch auction, to restore a good competition based on the fundamental analysis and not on the traders' arms race for speed.

Bibliography

Alfonso Puorro, 2013, High Frequency Trading: una panoramica, Questioni di Economia e Finanza n° 198, Banca d'Italia, 7.

Ananth Madhavan, 2000, Market Microstructure: A Survey, Journal of Financial Markets, Vol.3, issue 3, 205-258.

Andrei Kirilenko, Mehrdad Samadi, Albert S. Kyle, Tugkan Tuzun, 2017, The Flash Crash: The Impact of High Frequency Trading on an Electronic Market, The Journal of Finance, Vol. 72, No 3, 967-998.

Association for Financial Markets in Europe, The role of post-trade services in the financial sector, 2015.

Avanidhar Subrahmanyam, 2012, Stock Market Circuit Breakers, Government Office for Science.

Avanidhar Subrahmanyam, 2013, Borsa Istanbul Review, Volume 13 (3), 4-9.

Benn Steil, 2002, Changes in the Ownership and Governance of Securities Exchanges: Causes and Consequences, in Brookings-Wharton Papers on Financial Services, 61-91.

Blake F., 1986, Noise, The Journal of Finance, Vol. 41 (3), 529-543.

Blume Marshall E., Siegel Jeremy J., Rottenberd D., 1993, Revolution on Wall Street: the Rise and the Decline of the New York Stock Exchange, W.W. Norton & Company, New York.

Brogaard J., 2011. High frequency trading and market quality. Working Paper, University of Washington.

Brogaard, T. Hendershott, R. Riordan, 2013, High Frequency Trading and price discovery, European Central Bank Working Paper Series No. 1602.

CFTC (2015), CFTC Unanimously Approves Proposed Rule on Automated Trading, Press release.

CFTC (2015), Regulation AT.

Charles M. Jones, 2013, "What do we know about high frequency trading?", Columbia Business School.

Charles R.Korsmo, 2013, “High-Frequency Trading: A Regulatory Strategy”, U. Rich. L. Rev., Vol. 48,523-609.

Clara Martins Pereira, 2020, Unregulated Algorithmic Trading: Testing the Boundaries of the European Union Algorithmic Trading Regime, Journal of Financial Regulation, Vol. 6, 270-305.

Claudia E. Moise, Paca Flaherty, 2017, Limit Up-Limit Down Pilot Plan and Associated Events, U.S Security and Exchange Commission, white papers.

Code of Federal Regulation

Cyrille Guillaumie, Giuseppe Loiacono, Christian Winkler, Steffen Kern, 2020, Market impacts of circuit breakers – Evidence from EU trading venues, ESMA Working Paper No. 1.

Delegated Regulation 2017/578/EU.

Delegated Regulation 2017/578/EU.

Delegated Regulation 2017/584/EU

Delegated Regulation 2017/584/EU.

Delegated Regulation 2017/589/EU

Delegated Regulation 2017/589/EU.

Directive 2014/65/EU

Directive 2014/65/EU.

Donald E. Farrar, 1974, Toward a central Market System: Wall Street’s slow retreat into the future, The Journal of Financial and Quantitative Analysis, Vol. 9, No 5, 815-827.

Dondoni, A., Maggi, M., & Montagna, D., 2018 “VIX Index Strategies Shorting volatility as a portfolio enhancing strategy”. Universita’ degli Studi di Pavia, Banca IMI.

Edwin Hu, 2018, Intentional Access Delays, Market Quality, and Price Discovery: Evidence from IEX Becoming an Exchange, SEC Division of Economic and Risk Analysis’ Working Papers.

Ekkehart Boehmer, Kingsley Fong, Julie Wu, 2012, International evidence on algorithmic trading, Working Paper, EDHEC Business School.

Eric Budish, Peter Cramton, and John Shim, 2015, The high frequency trading arms race: frequent batch auctions as a market design response, Quaterly Journal of Economics, Vol.130 (4), 1547-1621.

ESMA, 2014, "High-frequency trading activity in EU equity markets", Economic Report, Number 1.

ESMA, 2017, Guidelines, Calibration of circuit breakers and publication of trading halts under MiFID II.

Eugene F. Fama, 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, The Journal of Finance, Vol. 25, No 2, 383-417.

Fox, Glosten, Gabriel, Rauterberg, 2015, The New Stock Market: Sense and Nonsense, Duke Law Journal, Vol. 65, No. 2, 191-277.

Gary Shorter, Rena S. Miller, 2014, High-Frequency Trading: Background, Concerns, and Regulatory Developments, Congressional Research Service, 23.

German High Frequency Trading Act. (2013).

Gomber P., Arndt B., Lutat T., Uhle T., 2011, High-Frequency Trading, Goethe Universitat Frankfurt Am Main

Guido Ferrarini, Paolo Saguato, 2014, Regulating Financial Market Infrastructures, ECGI Working Paper N°. 259.

Hagstromer, B. and Norden, L., 2013, The diversity of high frequency traders, Journal of Financial Markets, 16(4), 741-770.

Hagströmer, B., L. Nordén, and D. Zhang, 2014. How Aggressive Are High-Frequency Traders?, The Financial Review, Vol. 49(2), 395–419.

Harris L., 2003, Trading and Exchanges: Market microstructure for practitioners, Oxford University Press.

Hasbrouck, J. and Saar, G., 2013, Low-Latency Trading, Journal of Financial Markets, Vol. 16(4), 646-679.

Hasbrouck, J., and G. Saar, 2009. Technology and Liquidity Provision: The Blurring of Traditional Definitions. Journal of Financial Markets, vol. 12, issue 2, 143-172.

Hendershott T., Jones C. M., Menkveld A. J., 2011, Does Algorithmic Trading Improve Liquidity?, The Journal of Finance Vol. 66, No. 1, 1-33.

Hendrick Bessembinder, Kumar Venkataraman, 2009, Bid-Ask Spreads: Measuring Trade Execution Costs in Financial Markets, Encyclopedia of Quantitative Finance.

IOSCO, 2010, Issues raised by Dark Liquidity, Technical Committee of the IOSCO.

IOSCO, 2011, Regulatory Issues Raised by the Impact of Technological Changes on Market Integrity and Efficiency, Technical Committee of the IOSCO.

Irene Aldridge, 2010, High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems, John Wiley & Sons, Inc.

J. Angel, L. Harris, C.S. Spatt, 2010, Equity Trading in the 21st Century, USC Marshall School of Business.

J.C. Hull, 2017, Options, Futures and Other Derivatives, 10th ed., Pearson.

James L. Hamilton, 1987, Off-Board Trading of NYSE-Listed Stocks: The Effects of Deregulation and the National Market System, in the Journal of Finance, Vol 42, No. 5, 1331-1345.

Jared Egginton, Bonnie F. Van Ness, and Robert A. Van Ness, 2016, Quote Stuffing, Financial Management

Jerry W. Markham & Daniel J. Harty, 2008, For Whom the Bell Tolls: The Demise of Exchange Trading Floors and the Growth of ECNs, Journal of Corporation Law, Vol. 33, No 4, 865-939.

Johnson B., 2010, Algorithmic Trading and DMA: An Introduction to Direct Access Trading.

Law Decree of the Italian Minister of Economy and Finance of 21 February 2013.

Lawrence E. Harris, Ethan Namvar, 2016, The Economics of Flash Orders and Trading, Journal Of Investment Management, Vol. 14, No. 4, 1-17.

Lupoi A., 2020, La struttura del mercato ed i riflessi giuridici, CEDAM.

Macey, O'Hara, 2005, Markets to venues: Securities regulation in an evolving world, in Standford Law Review, Vol 58, No 2, 563-600.

Maria Coelho, 2016, Dodging Robin Hood: Responses to France and Italy's Financial Transaction Taxes, University of California, Berkeley, College of Letters & Science, Department of Economics

Martin Haferkorn, Kai Zimmermann, 2014, The German High-Frequency Trading Act: Implications for Market Quality, Goethe University Frankfurt.

Martin W. McChesney, 1971, The Securities Markets: A report with Recommendations, in Virginia Law Review, vol 51, No 7.

Menkveld A., 2013, High frequency trading and the new-market makers, Journal of Financial markets, Vol 16(4), 712-740.

Michael Morelli, 2017, Implementing High Frequency Trading Regulation: A Critical Analysis of Current Reforms, Michigan Business & Entrepreneurial Law Review, Vol.6 (2), 201-229.

Office of Analytics and Research, Division of Trading and Markets, SEC, Empirical Analysis of Liquidity Demographics and Market Quality for Thinly-Traded NMS, April 2018.

Open Letter, BATS, Market Structure Reform Discussion, at 3-4 (Jan. 6, 2015).

Poblocka, Iwona, 2019, High Frequency Trading: highway to financial hell or economic salvation : a comprehensive review of the High Frequency Trading literature, University of Twente, Industrial Engineering and Management MSc.

Rebecca Lewis, 2015, A new approach to stock market execution, The Federal Reserve Bank of Chicago, Chicago Fed Letter, Number 343.

Regulation 596/2014/EU.

Regulation 600/2014/EU.

Regulation 648/2012/EU.

Report issued by Merryl Lynch Pierce Fenner & Smith Inc. on june 14 1979 to US subcommittee, Progress toward the development of a National Market System.

Report of the Staffs of The CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues, September 30, 2010, Findings Regarding the Market Events of May 6, 2010.

Ronald Gilson & Reinier R. Kraakman, 1984, The Mechanisms of Market Efficiency, Virginia Law Review, Vol. 70, No. 4, 549-644.

SEC (1998), Regulation of Exchanges and Alternative Trading Systems, Release No. 34–40760.

SEC (1999), Regulation ATS, Release No. 34-40760.

SEC (2005), Regulation NMS, Release No. 34-51808.

SEC (2010), Concept Release on Equity Market Structure, Release No. 34-61358.

SEC (2010), Concept Release on Equity Market Structure, Release No. 34-61358.

SEC (2012), Investor Bulletin: Measures to Address Market Volatility, Investors Alerts and Bulletins.

SEC (2012), SEC Approves Proposals to Address Extraordinary Volatility in Individual Stocks and Broader Stock Market, Press Release.

SEC (2013), SEC charges Knight Capital with violations of Market Access Rule, Press release.

SEC (2013), Trade Execution: What Every Investor Should Know, Investor Publications.

SEC (2015), SEC Proposes Rule to Require Broker-Dealers Active in Off-Exchange Market to Become Members of National Securities Association, Press release.

SEC (2016), Investor Alert: Tick Size Pilot Program – What Investors Need To Know, Investors Alerts and Bulletins.

SEC (2020), Report by the Staff of the U.S. Securities and Exchange Commission on Algorithmic Trading in U.S. Capital Markets.

Stijn Claessens, 2019, Fragmentation in global financial markets: good or bad for financial stability?, BIS working papers.

Sullivan & Cromwell LLP, 2012, SEC Order Approves Proposals to Address Extraordinary Volatility in Individual Stocks and Broader U.S. Stock Market.

The Nasdaq Stock Market LLC, Notice of acceptance Acceptance, Waiver and Consent (AWC), June 16, 2014.

Tobias Ruhl & Michael Stein, 2014, The Impact of Financial Transaction Taxes: Evidence from Italy, 34 Economic Bulletin 25, 32.

U.S. Equity Market Structure: An Investor Perspective, VIEWPOINT (BlackRock, New York, N.Y.). Apr. 2014.

Valeria Caivano, 2015, The impact of High-Frequency Trading on Volatility: Evidence from the Italian Market, Quaderni di Finanza n° 80, Consob.

Valeria Caivano, Salvatore Ciccarelli and Giovanna Di Stefano, 2012, Il Trading ad Alta Frequenza: Caratteristiche, Effetti, Questioni di Policy, CONSOB Discussion Papers No. 5, 22.

William G. Christie, Paul H. Schultz, 1994, Why Do NASDAQ Market Makers Avoid Odd-Eighth Quotes?, Journal of Finance, Vol. 48, No 5, 1813-1840.

Yesha Yadav, 2015, How Algorithmic Trading undermines efficiency in capital markets, Vanderbilt Law Review, Vol. 68 (6), 1607-1671.

Yesha Yadav, 2016, Structural insider dealing, UCLA Law Review, Vol. 60, 968-1033.

Yeshav Yadav, 2013, The Problematic Case of Clearinghouses in Complex Markets, Georgetown Law Journal, Vol. 101, 387- 444.

Zhang, Sarah and Ryan Riordan, 2011, Technology and market quality: the case of high frequency trading, ECIS 2011 Proceedings, Paper 95.

Sitography

Bloomberg: www.bloomberg.com

Commodity Futures Trading Commission: www.cftc.gov

Financial Industry Regulatory Authority: www.finra.org

Financial Times: www.ft.com

Nanex: www.nanex.net

New York Stock Exchange: www.nyse.com

The Wall Street Journal: www.wsj.com

U.S. Securities and Exchange Commission: www.sec.gov

Unlisted Trading Privileges: www.utpplan.com

US Government Publishing Office. www.gpo.gov