University of Hertfordshire

HIGH FREQUENCY TRADING: A THREAT TO FINANCIAL STABILITY?

Ph.D. Candidate:	Gianluca Virgilio
Principal Supervisor:	Dr. Georgios Katechos
Second Supervisor:	Dr. Maria Schilstra
Second Supervisor:	Dr. Aarti Rughoo
Date:	1 February 2017

Submitted to the University of Hertfordshire in partial fulfilment to the requirements of the degree of PhD

ABSTRACT

The purpose of this thesis is: (i) to produce an in-depth data analysis and computer-based simulations of the market environment to investigate whether financial stability is affected by the presence of High-Frequency investors; (ii) to verify how High-Frequency Trading and financial stability interact with each other under non-linear conditions; (iii) whether non-illicit behaviours can still lead to potentially destabilising effects; (iv) to provide quantitative support to the theses, either from the audit trail data or resulting from simulations. Simulations are provided to test whether High-Frequency Trading: (a) has an impact on market volatility, (b) leads to market splitting into two tiers; (c) takes the lion's share of arbitrage opportunities. Audit trail data is analysed to verify some hypotheses on the dynamics of the Flash Crash.

The simulation on the impact of High-Frequency Trading on market volatility confirms that when markets are under stress, High-Frequency Trading may cause volatility to significantly increase. However, as the number of ultra-fast participants increases, this phenomenon tends to disappear and volatility realigns to its standard values.

The market tiering simulation suggests that High-Frequency traders have some tendency to deal with each other, and that causes Low-Frequency traders also to deal with other slow traders, albeit at a lesser extent. This is also a kind of market instability.

High-Frequency Trading potentially allows a few fast traders to grab all the arbitrage-led profits, so falsifying the Efficient Market Hypothesis. This phenomenon may disappear as more High-Frequency traders enter the competition, leading to declining profits. Yet, the whole matter seems a dispute for abnormal gains only between few sub-second traders.

All simulations have been carefully designed to provide robust results: the behaviours simulated have been drawn from existing literature and the simplifying assumptions have been kept to a

minimum. This maximises the reliability of the results and minimizes the potential of bias.

Finally, from the data analysis, the impact of High-Frequency Trading on the Flash Crash seems significant; other sudden crashes occurred since, and more can be expected over the next future.

Overall, it can be concluded that High-Frequency Trading shows some controversial aspects impacting on financial stability. The results are at a certain extent confirmed by the audit trail data analysis, although only indirectly, since the details allowing the match between High-Frequency traders and their behaviour are confidential and not publicly available Nevertheless, the findings about HFT-induced volatility, market segmentation and sub-optimal market efficiency, albeit not definitive, suggest that careful monitoring by regulators and policy-makers might be required.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I wish to thank my supervisors at the University of Hertfordshire, Dr. Georgios Katechos, Senior Lecturer in Finance, Dr. Maria Schilstra, Senior Lecturer in Computer Science, and Dr. Aarti Rughoo, Senior Lecturer in Economics, for their help and useful advices on this work.

For the completion of this work I am also indebted with several other people; without their help this research could not even be initiated.

Only exception to the alphabetical order used below, and first person I wish to thank, is Dr. Francesca Campolongo, Head of the Finance & Economy Unit at the European Commission's Joint Research Centre, for her continuous support.

Dr. Carlo Ferigato, researcher at the Joint Research Centre, invaluable for his deep knowledge about Petri Nets, and not only.

Dr. Sven Langedijk, Head of the Modelling, Indicators & Impact Evaluation Unit at the Joint Research Centre and local supervisor, his suggestions provided a light in the darkest moments of my effort.

Dr. Pierpaolo Malinverni, Programme Manager at the Joint Research Centre: his encouragement proved invaluable.

Dr. Andrea Pagano, researcher at the Joint Research Centre, a reference person for all doubts and questions I had throughout these years.

Dr. Domenico Perrotta, researcher at the Joint Research Centre and local supervisor, a real support for, and beyond, any advice I needed.

All the people mentioned were generous in advising and suggesting me the way forward.

Any errors, weaknesses and deficiencies of this research are exclusively my own and sole responsibility.

FOREWORD

In order to avoid any gender issues, and to make the reading smoother, in the following, traders, brokers, investors, market participants, agents, players, operators, and the like, will be intended to be firms rather than individuals. In such a way, they will be referred to by using the neutral pronoun 'it' and the possessive 'its', rather than the androgen pronouns 'he/she', 'him/her', and the possessive adjectives 'his/her'.

ACRONYMS

ABM	Agent-Based Model
ANOVA	Analysis Of Variance
AT	Algorithmic Trading
BATS	Better Alternative Trading System
CFTC	Commodity Futures Trading Commission
CME	Chicago Mercantile Exchange
CQS	Consolidated Quotation System
CTS	Consolidated Trading System
DBMS	Data Base Management System
DJIA	Dow Jones Industrial Average
EDT	Eastern Daylight Time
EMH	Efficient Market Hypothesis
ETF	Exchange Traded Fund
FIFO	First In-First Out
FPGA	Field Programmable Gate Array

FX	Foreign eXchange
GMT	Greenwich Mean Time
HF	High Frequency
HFT	High-Frequency Trading
ISO	Intermarket Sweep Order
LCTM	Long-Term Capital Management
LF	Low Frequency
LFT	Low-Frequency Trading
LOP	Law of One Price
LRP	Liquidity Replenishment Point
LSE	London Stock Exchange
MiFID	Markets in Financial Instruments Directive
NASDAQ	National Association of Securities Dealers Automated Quotation
NBBO	National Best Bid Offer
NBER	National Bureau of Economic Research

NYSE	New York Stock Exchange
OTR	Order-Trade Ratio
PN	Petri Net
QR	Quoting Ratio
RegNMS	Regulation National Market System
RSS	Real Simple Syndication
S&P	Standard & Poor
SEC	Securities and Exchange Commission
SIP	Securities Information Processor
SL	Stop Loss
TA	Technical Analysis
TAQ	Trade And Quote
VIX	Volatility IndeX
VPIN	Volume synchronized Probability of Informed trading
W&R	Waddell & Reed

1. INTRODUCTION

Digitalisation of the financial markets is one of the main innovations that appeared in the financial arena over the last few decades. Although in a poorly coordinated manner, most exchanges around the world have migrated from the old-fashioned open outcry environment to electronic trading systems, and today this way of operating is by far the most predominant.

The consequences of this epochal change are far-reaching and in some cases even dramatic.

After a few years since market digitalisation a completely new way of trading has appeared. Orders are transmitted over the wires rather than face-to-face or by telephone, and this fact paves the way to new strategies aimed at exploiting the incomparable speed of operations computers enjoy facing humans. Speed has more and more become the major competitive advantage in the trading environment. In a first-come-first-served environment, being able to quote a limit order before other (slower) traders is a guarantee of more likely execution. Financial, commercial, business and political news need to be interpreted before competitors do in order to exploit the chance to quote, cancel and re-quote orders, hence minimising the risk of being picked off by more informed investors. Small price swings need to be noticed and acted upon quickly if arbitrage opportunities or minimal price fluctuations are to be exploited. Even market manipulation becomes easier to bring in and harder to spot out. Regulators and policy makers find themselves far behind in this sort of arms race, where expenditure in hardware, software and networking now makes up for the largest part of many financial firms' budgets.

Such dramatic changes rise new concerns about several aspects most people are not aware of, yet.

The nature of the problem being investigated in this research is that High-Frequency Trading (HFT) has been regarded as a possible threat to the 'fair' game slower investors were used to, a source of financial instability, a denial of the historic purpose market had in conveying resources to the most promising companies, a vehicle for market manipulation and, above all, a generator of more frequent crashes and bubbles.

Professional traders not directly involved in HFT have been critical since the very beginning of this practice [Beunza, Millo and Pardo-Guerra (2012)]. Quite the contrary, the academic world, more hard fact-orientated than practitioners, struggles to find evidence of HFT's negative aspects. Albeit with some contrasting voices, many scholars looked at the positive side of HFT, as higher liquidity leading to lesser price impact of large transactions [Aitken et al. (2012), Barker and Pomeranets (2011)], lower transaction costs facilitating widespread participation [Foucault, Kadan and Kandel (2013), Kirilenko and Lo (2013)], quicker price discovery accelerating convergence toward the price suggested by fundamentals [Brogaard, Hendershott and Riordan (2014), Manahov and Hudson (2014)]; reduced volatility enhancing investors' confidence in the orderly functioning of the financial markets [Hasbrouck and Saar (2013), Myers and Gerig (2014)], and thinner bid-ask spread with potentially higher profit also for occasional investors [Zervoudakis et al. (2012), Menkveld (2013)]. The net result is usually interpreted to be a clear improvement of market quality with benefits for all investors, irrespective of their starting capital, adopted strategy, and quoting or trading speed. These are the mainstream findings, but not all studies agree with this view. Some authors notice that volatility may be generally lower, but when it turns up its effects are much more significant [Kirilenko et al. (2011), Jarrow and Protter (2012)]. High-Frequency traders may tend to interact more often than expected with their similar counterparts, leading to splitting the market in two lanes with the traders' speed as the main discriminant [Cartlidge and Cliff (2012), Hasbrouck and Saar (2013)]. Price discovery is certainly faster, but perhaps only to the benefit of the ultra-fast traders, who grab the large majority of the short-term opportunities [Aldridge (2010), Lewis (2014)]. New metrics are being produced to shed light on the implicit danger of market toxicity created by the enormous speed difference between fast and slow traders [Easley, Lopez de Prado and O'Hara (2011)].

A large number of academic papers on the topic has been published in the last few years: a Google Scholar search on 'High Frequency Trading' in early 2017 found more than one million results.

Despite such a large amount of literature the answer to the basic question "*is High-Frequency Trading beneficial to the financial markets and on financial stability in general?*" seems not to have been reached yet. The sharp differences among academics' opinions and between practitioners and academics highlight the need for a deeper investigation, providing further insights on this issue. Therefore, further research seems appropriate.

The aims of this research can be summarised as follows:

1. Examine the impact of HFT on price volatility by using computer simulations to verify whether market activities of a small number of ultra-fast traders changing prices very rapidly can make slower quotes obsolete between the time they are conceived and the time they hit the books. If so, a market order may transact at a price different from the one originally intended, increasing volatility.
2. Examine the possibility of markets splitting into a fast and a slow lane with little interaction between them. I use computer simulations to test whether, as hypothesised by Hasbrouck and Saar (2013), algorithms which submit and cancel quotes in tiny orders of magnitude, unreachable by traditional traders (milliseconds or below), target their similarly fast fellows, implicitly setting the framework of two separate trading lanes.
3. Investigate the effect of HFT on the validity of the Efficient Market Hypothesis. Again I employ computer simulations to reproduce an environment in which a small number of ultra-fast traders compete with the rest of the (much slower) operators to grab arbitrage opportunities. The objective is to verify whether there is a statistically significant difference between the amounts of arbitrage opportunities exploited by the two communities.
4. Analysis of factors that likely contributed to one of the most dramatic event allegedly attributable to HFT (the May 6, 2010 Flash Crash): The objective aimed at by the analysis, based on audit trail data, is the evaluation of the impact HFT and stop-loss orders had on price volatility,

5. Another objective, to be achieved through both a theoretical model and audit trail data analysis, is to verify whether high-frequency quote cancellations can, and did, lead to market anomalies.

The analyses have HTF impact on financial stability as a reference and its contribution to market quality as a purpose. The methodological framework used throughout this research is based on computer simulations employing Agent-Based Models; a mathematical model (Petri Nets) reproducing some specific behaviours of High-Frequency traders; data analysis of the Flash Crash comparing its main parameters with those of the other days in the same week.

The following paragraphs of this chapter introduce the subsequent chapters of the thesis. In particular, the structure of the research is explained in paragraph 1.7, whereas sections 1.7.3, 1.7.4, 1.7.5, and 1.7.6 describe with the level of details suitable for an introductory chapter the objectives of each of the four main branches of the thesis. Chapter 3 explains in deeper details the aims and objectives of the research: the impact of HFT or volatility (3.1.3), market tiering (3.1.4), HFT and EMH (3.1.5) data analysis (3.1.6) and the issue about relative vs. absolute speed (3.1.7) Moreover, it explains the rationale behind using simulation as a research tool.

1.1. THE PILLARS: FINANCIAL STABILITY AND HIGH-FREQUENCY TRADING

When discussing financial stability, it is commonly accepted that economic and financial aspects are tightly intertwined. It is still a matter of debate today whether the 1929 Great Crash originated within the stock exchange system and propagated to the real economy or a bubble in the latter triggered speculation and subsequent disastrous fall in the former. The same holds for the crashes of 1987 and 2001 (as well as many others). Supposedly economic-side crises, like the one originated in the subprime sector (2007-2008), can be regarded instead as triggered by financial speculation protracted far too long. But even though the original supposition should prove correct, the crisis had a profound impact on the finance industry, with major financial institutions going bankrupt or requiring exceptional intervention from their governments to be rescued - at the expenses of the taxpayer, something which made the crisis of the real economy even more pronounced. Given its scope, this research shall investigate solely the aspects of financial stability affecting the financial markets and not the economic scenario at large.

Since the late Eighties a new way of financial trading began: exchanges started to deal with transaction electronically and participants started to convey their order messages via the same technology. A profound change had occurred. High-Frequency Trading is the extreme, yet logical, consequence of the introduction of computers and data networks on financial markets. As long as financial transactions were carried out in an open outcry trading floor the risks were, albeit with some notable exceptions, mainly confined to the realm of fundamentals. Wars were a definitive cause of stock prices going abruptly up or down; earthquakes, flooding, droughts, and other acts of God used to drive them. Failure to grant agreements on profitable contracts or to hit revenue targets were other more mundane sources of sudden stock price drops. That was true then – and still is today. Yet, since about three decades, when transactions started to be transmitted over wires, the world changed – and forever. The open outcry is increasingly an exception and the computer the rule. When we think of a financial transaction, no longer do we figure out a hand-shake but rather an electronic impulse or a mouse click. With a little imagination today's world could have been foreseen in the mid-Eighties, when the human to electronic transition started in many stock exchanges. What was definitely harder to imagine was that over the last ten years the order resting time could drop from an order of magnitude of minutes down to seconds and then to milli-, or perhaps, micro-seconds. This is the essence of High-Frequency Trading. Research [for example Beunza, Millo and Pardo-Guerra (2012)] has shown that to most human practitioners HFT immediately reminds the sudden price plunge (and nearly as quick recovery) which occurred on May 6, 2010, the so-called Flash Crash. No fundamental news, at least not able to move the market down that percentage. No time to reflect or even to understand what was happening. No unique explanation; too many and too mutually contrasting for being unanimously accepted. Only the feeling that the world had lost one trillion dollars in less than six-and-a-half minutes, or more than two and a half billion dollars per second, for three hundred and eighty-eight interminable seconds. And that it could happen again, at any time. May be next month, tomorrow or on the next microsecond. This is the essence – and the big worry - of the Flash Crash.

The following paragraphs address the topics mentioned in the title of this research (High-Frequency Trading and financial stability) with the purpose of clearly define their meaning and using them in the rest of the thesis. Clear definitions are necessary because, as it will be seen, they are not so straight-forward, nor unambiguous. Paragraph 1.2 presents a definition of financial stability, then the closely related concept of risk (in particular, endogenous risk) is discussed (paragraph 1.3) and the financial stability parameters most relevant to trading are illustrated, namely market efficiency, price discovery, volatility, liquidity, bid-ask spread, transaction costs, and fairness to all. Then, it follows paragraph '1.5. Definition of High-Frequency Trading' and the relation HFT has with the financial markets (paragraph 1.6). The paragraph 1.7 presents the structure of the research and the last paragraph concludes.

1.2. DEFINITION OF FINANCIAL STABILITY

Despite having been the focus on financial stability issues at the forefront since a long time, and especially in the aftermath of the subprime crisis, still the financial community has been unable to converge over a generally accepted definition of 'financial stability'. Different backgrounds, operating conditions, observation points, interpretations, opinions, sometimes lobbying interests and temporary contingencies are at the root of this incapacity to agree upon a universally-accepted definition. The matter is made even more complex by the deep gap existing between the real, hard economy and the digital, virtualised financial sector: the meaning of financial stability for one seems to be different from the others. Fell and Schinasi (2005) identified three main characteristics of financial stability: (i) the efficiency with which the financial system facilitates the allocation of resources to businesses; (ii) the sensible assessment and good management of financial risk through accurate pricing of financial assets; and (iii) the smoothness with which markets and participants absorb shocks. Shocks are indeed regarded as a major threat to financial stability, as the former can trigger the latter and vice-versa in a sort of self-reinforcing feedback loop. The main effect of shocks is the degraded capability of the markets to accomplish their main function: channelling funds to promising investment opportunities. Protracted malfunctioning of the markets may, and often does, adversely affect the real economy. Economy and finance are tightly intertwined. When asset prices substantially diverge from their fundamental value, like a tsunami the long wave floods into, and causes, the reduced credit availability to families and businesses, which can no longer honour their obligations, often with dramatic consequences. An operational attempt to define the concept is made by Alawode and Al Sadek (2008), who indirectly define financial stability as requiring “that the key institutions in the financial system are stable, in that there is a high degree of confidence that they continue to meet their contractual obligations without interruption or outside assistance; and that the key markets are stable, in that participants can confidently transact in them at prices that reflect the fundamental forces and do not vary substantially over short periods when there have been no changes in the fundamentals” (ibid. p.9). It seems obvious that, because of the impact on everybody's life, the matter of the debate is by no means just theoretical; it extends to very operational aspects, like the ones dealt with by those institutions which have the financial stability in their very mission. The two authors present a number of definitions taken by the official documents of some national Central Banks. However, different institutions show different viewpoints, often driven by the specific situation experienced in their respective countries. So, while the Deutsche Bundesbank links financial stability to the key macroeconomic functions and the Reserve Bank of Australia aims to keep the financial system stable to help promoting economic growth, the Bank of Japan seems mainly worried in preserving firms' and individuals' confidence in the system. The Austrian National Bank stresses the need of the markets to keep working satisfactorily even in the case of shocks whereas the Central Bank of Argentina explicitly mentions the sustainability of a nationwide payment system and the difficulties experienced by the population who saw its savings evaporated. A similar concern is expressed by another institution that recently faced a dramatic crisis, the Central Bank of Iceland, which praises a stable system suitable to withstand shocks to the economy and financial markets whereas the less worried Swiss National Bank only expects intermediaries and infrastructures to fulfil their respective functions and prove resistant to shocks, just regarded as 'potential'. It is understandable that Central Banks focus attention on the aspects of financial stability more relevant to the history of their own country but it must be noticed how such differences, although apparently minor, when translated into practical behaviour and regulations impact the financial environment in one way or another - and it is difficult to state that national policies are well coordinated.

Given the difficulty of reaching a common definition of financial stability, a possible solution has been found in its opposite: defining financial instability. In this respect things seem clearer. Answers range from the "conditions in financial markets that harm or threaten to harm an economy’s performance" [Alawode and Al Sadek (2008), p. 8] to the link "between the volatility in asset prices and the consequent flows through to the real economy" [Foot (2003), no pagination], to the definitive description Walter Bagehot, a distinguished XIX century journalist, made in the 1870s about financial crisis, identified as "when the Bank of England is the only institution in which people have confidence" [cited in Foot (2003), no pagination]. Indeed, financial instability is often viewed as a prerequisite for a crisis, even if it must be noticed that the outcome may well range from a stock market collapse to a banking crisis, from oil price shocks to real estate quakes. However, what most authors agree about is the continuum along which financial instability move [Fell and Schinasi (2005)], or the corridor financial systems operate within [Alawode and Al Sadek (2008)]. Also according to Foot (2003) a threat to financial stability has to be measured against the rapidity with which asset prices rise or fall, implying that a continuous variable (rapidity) will give rise to a continuous phenomenon: "[b]ut, at some point, the relevant asset prices will rise or fall so fast that what was a local phenomenon starts to affect the real economy and thus potentially to affect financial stability" (#23). Proof of non-existence of a stability-instability threshold is that several (relatively) limited crises, as occurred in the (relatively) quiet Nineties, caused more bank failures than in the roaring Thirties (#32).

1.3. ENDOGENOUS RISK

Financial stability, or lack of it, directly leads to the concept of risk. First of all, it must be clear that risk is a useful feature of financial markets. Risk encompasses the possibility of incur into a loss as well as to make large profits. Financial markets would not even exist if no risk was entailed. Risk level also runs on a continuous slide, from minimum to maximum, without holes or discontinuities in-between. But in order to identify what is a 'reasonable' risk level, which allows financial activity without allowing instability, the term is to be defined. Once again, there seems to be no unique definition of the concept. Danielsson (2013) recognises multiple meanings. A common textbook definition would state it in terms of volatility of returns, two prerequisites being the normal distribution of the returns and a constant volatility. In the real world neither assumption holds. Volatility of returns can and does vary as a complex function of a very large number of parameters, including intuition and interpretation of the future based on present conditions. Under no simplifying assumption can volatility be considered constant - nor need returns to be normally distributed. Laws governing risk are inherently non-linear and the more financial systems (being markets, institutions, participants, beliefs, expectations, mathematical formulae and greed all part of the 'system') are interconnected, and the less protection from contagion and the riskier the system is. Fell and Schinasi (2005) produced a list of all possible risk factors. Among others they identified credit risk, market risk, liquidity, interest rate, currency, operational risk, information technology weaknesses, legal/integrity risk, reputation, business strategy, concentration, capital adequacy, counterparty risk, price misalignments, contagion, clearance, payment and settlement risk, infrastructure fragility, collapse of confidence, domino effects. And then fully-exogenous risks, like natural disasters, political events, large business failures. Although not explicitly mentioned therein, diplomatic crises, wars and terrorism can also be added. However, it must be stressed that, although the exogenous risk sounds as the most worrisome, it is actually the endogenous risk that most often put the financial system in jeopardy. Danielsson (2013) is rather clear about it: "[e]ndogenous risk arises when individual economic agents react to their environment and their actions in turn affect their environment to such a degree that an endogenous feedback cannot be ignored. Financial markets, where all market participants are constantly competing against each other, trying to gain advantage by anticipating each other’s moves, are a clear example of an environment creating endogenous risk" (ibid. p.55). Usually endogenous risk is low because most agents interacting within a system behave independently from one another, and so doing balance each other out. However, it does occasionally happen that this equilibrium status breaks off - and the consequences may be disastrous. The Millennium Bridge is brought as an example. For reasons that will be explained at length in chapter 5, the interaction between a gentle Thames breeze and the pedestrians walking over the bridge moving in sync as a reaction, caused large and self-reinforcing bridge oscillations, leading to widespread panic. Indeed, the originating factor could well have been a different one and still the outcome be the same. The financial markets show an astonishingly similar behaviour. A small shock randomly appears, and then many participants start moving in sync like pedestrians on the bridge. If a sufficiently large number of them trade in the same direction, their behaviour will create a feedback on the markets causing even more traders to join the flow. As noticed by Danielsson (2013), "[e]ndogenous risk tends to be low most of the time, because economic agents usually behave individually and have different objectives and information sets. This means that in aggregate their behaviour resembles noise when viewed from the outside. Under certain conditions, market participants start behaving much more harmoniously than usual, amplifying price movements that result in asset price bubbles and finally market crashes" (ibid.

p.55). Of course, not all price movements cause a major crisis, like no all breezes wobble a well-engineered bridge. But sometimes both kinds of event occur - and financial markets, much more often than bridges, resonate with their users.

1.4. MARKET PARAMETERS

There are several parameters relevant to financial stability of the markets. The main ones are briefly introduced in the following.

1.4.1. MARKET EFFICIENCY

In order to play a fair game the outcome must not be known in advance [Samuelson (1965)]. Although this may sound an obvious, and even trivial, statement, most market participants devote a lot of time and effort to falsify it. Search for alpha (excess return above the one compatible with the chosen risk level) is the goal of all professional traders. Yet, theoretical findings go into the opposite direction: the Random Walk is a comfortable certainty in universities; not so on trading floors. Practitioners can belong to the fundamental or to the technical analysis party, they may be quants or intuitionist, some trade on private information or claim using superior algorithms but none of them believe that prices, as stated by Fama (1965), follow a random walk. Otherwise they would not be trading for living. This may be depicted like the 'trader's paradox': traders like financial stability to implement their strategies with no disturbances but the very concept of strategy imply that prices can be foreseen, something which would drastically lean toward financial instability.

1.4.2. QUICK PRICE DISCOVERY

Asset price must be right. And the sooner it gets so and the better. Mistaken prices create noise, confusion and lead to potential instability. On the other side, if asset prices were always right all the time, nobody would trade. There is no point in paying the right price for an asset. The seller may not know but the buyer values the asset she bought more than its price. And similarly in the seller's view the asset price is higher than its worth. Since both parties are confident in their view, they both hurry to close a deal they deem profitable. The price may actually not be right but speed of execution is. If a sufficient number of participants evaluate the asset worth rightly, the price, may be after a few oscillations, stabilises around its fair value. If this process takes too long, it is likely to create uncertainty about the true price and instability might ensue. Quick price discovery allows analysts to concentrate on the future value of the asset rather than struggling to understand why its price did not aligned with the current value.

1.4.3. REASONABLE VOLATILITY

A certain level of volatility must exist in order to have a market. As long as currencies were pegged to a fixed value no currency market existed. But as soon as national devises were allowed to float freely, currency risk was born and a foreign currency market started flourishing. Danielsson (2013) argues that, despite what classical theory states, the actual risk of an asset is not necessarily equivalent to volatility of its price: it is the risk perceived by the market participants that may cause sharp and sudden volatility. When prices are rising steadily, market participants get confident about the rosy future - and buy. This wave of purchases drives prices up, which spreads the bullish sentiment, leading to more buys, and so on. Perceived risk is low but actual risk mounts. In the extreme cases this may develop into a bubble but even when the price rise is controlled (and so is volatility), the risk increases nevertheless. Self-defined wise investors celebrate easy earnings and remain oblivious about rising risk. When eventually the courses invert their direction, prices fall, sometimes abruptly, displaying high volatility and prices at, or below, book value. At that point most participants perceive the financial market as risky, but that is another illusory perception. Low volatility does not necessarily mean low actual risk but usually means low perceived risk - and vice versa. Regulators work hard to prevent uncontrolled volatility, as that often speaks financial instability. As Anderson et al. (2015) put it, "although episodes of heightened volatility and short-term illiquidity are not necessarily in themselves threats to financial stability, they could become so if they were to persist, amplify or spill over" (ibid. p.4). Moderate volatility all the time is the common goal of (non-speculating) investors, exchanges and policy makers alike.

1.4.4. SUFFICIENT LIQUIDITY

Liquidity is a sought-after characteristic of financial markets, as it represents the ease of converting assets into cash. Since rational investors are looking for liquid assets, markets being regarded as liquid attract clients - and their fees. In recent years some new entrants have designed their fee structure so to be more attractive for liquidity suppliers. Deeper books entice large market orders, since they have a lesser impact on price, and large orders translate into large fees. A flat-liquidity market in which order execution adversely affects the price would be regarded as rather poor and big clients posting large orders would prefer to stay away from it. Ideally, prices should be driven by fundamental, technical or other kind of analysis, not by order execution. Sophisticated traders deploy techniques for minimising the impact of their orders on price but if liquidity is scarce a certain percentage of a large order would still be executed at a loss compared to the price ex-ante. On the other hand, fee-capturing is not the only reason for promoting deep liquidity. According to some lines of thought, market liquidity was less than properly appreciated prior to the subprime crisis (2008), suggesting a link of the latter to the former. That was not the only case however, another example among many, being the autumn 1998 crisis, when liquidity in most mature world markets suddenly disappeared. Fat liquidity may be capable of absorbing shocks, to a certain extent, and therefore can be a factor of market stability. Indeed, in the aftermath of the subprime crisis, liquidity provision has become a high priority to policy makers.

1.4.5. THIN BID-ASK SPREAD

The bid-ask spread represents the compensation earned by the market makers against their obligation to provide liquidity to the market and for the risk of being picked off by more informed investors. When the tick (minimum price quantum) in the US was reduced from 1/8 of a dollar down to one centime, the scenario changed substantially. Market makers saw the potential profit per transaction to drop by more than twelve times and the number of transactions to rise. How this change affected financial stability is still a matter of debate but it is beyond doubt that it favoured higher frequency of transactions. Those traders that enjoy high speed of access to the book and that content themselves of tiny profit per individual transaction would thrive in the new environment. At the same time, high speed might be exploited to minimise the risk of being adversely selected by a more informed trader just by reducing order's resting time to a few seconds or less. The impact of this new way of operating on financial stability is unclear. On one side it made the books more densely populated, that is, more liquid. This seems to go in the direction of augmenting financial stability, whereas on the other side the increased order frequency, favoured by the closer price levels, made control more difficult. Because of sub-second trading, mistaken orders may no longer be recognised as such by humans before the damage has been done and software bugs may have far-reaching consequences. Moreover, the new scenario enticed many High-Frequency (HF) traders to supply liquidity, and enjoy the bid-ask spread, as market makers do but without the obligations the latter have. The main obligation is to supply liquidity under any market conditions, which may account to little when conditions are good but requires a lot of stamina when markets are under stress. So the new entrants seem to get the better of both worlds: they make profit, albeit tiny, per each transaction because they gain the spread, they minimise risk of being picked off thanks to the high-frequency resting order cancellations, and may withdraw without penalties when the going gets tough. Voices have been raised against this state of affairs but no definitive conclusion about the impact of tight spread on financial stability has been reached so far.

1.4.6. MODERATE TRANSACTION COSTS

The expression 'transaction costs' includes all the cost of a trade, the most substantial usually being bid-ask spread and commissions. As seen in the previous section, decimalisation of the tick contributed to reduce bid-ask spread and related costs.

Exchanges charge commissions to traders but since they are subject to marketing forces as any other kind of business, their fee structure changed as the environment did. It is in the interest of an exchange to have as many clients, i.e. traders, as possible operating on their premises, either physically or electronically. With the entrance of new trading venues, in order to attract clients, exchanges had to adapt their fee structure to the typology of their target customer segment. In the last few years traditional exchanges lost market share to the benefit of new entrants ready to grant liquidity suppliers a rebate, or to levy lower fees per transaction in view of a higher number of transactions per day. Reduction in transaction costs spells higher profits for market participants and a more diffuse participation enhances market attainment of its main goal: directing funds to the most promising projects, which in turn pushes economic growth for (good) businesses and wealth for the (good) investors. Therefore, according to classical theory, reduction in transaction costs works in the direction of wider participation and higher financial stability. On the other side reduced costs lead to more frequent access to the markets, and it did so in a way that is now difficult to keep under control by policy makers, regulators and exchanges alike. And lack of control certainly does not go in the direction of improved financial stability.

1.4.7. FAIRNESS TO ALL PARTICIPANTS

The financial market, as demonstrated by Samuelson (1965), is supposed to be assimilated to a fair game. In his article, the 1970 Nobel Prize winner states the Theorem of Fair-Game Futures Pricing that implies “there is no way of making an expected profit by extrapolating past changes […] by chart […] or mathematics” (ibid. p.44). This theorem is paramount for attracting to the markets not only the main institutional investors and the large banks but also the myriad of small financial companies and the individual day traders, down to the very occasional savers. All these participants must be ensured a fair treatment by the theory and practice and, above all, equal rules for everyone. Yet, in recent years controversial phenomena appeared, namely co-location and preferential convey of tape data. Co-location is a service offered by some exchanges to host investors' computers at the exchange premises, for a fee, in order to cut down networking latency. Another service is the contemporaneous transmission of tape data to the official server (which will distribute them to the public) and to investors' servers (again, for a fee). This way the regulatory requirement of providing data at the same time to all participants is met but it is also obvious that some investors will receive information before others. This issue has been barely raised so far but it seems hardly deniable that it creates all men unequal. In a business where information is at the very core, time unbalances risk to jeopardise confidence in the functioning of the market. And a market which lacks confidence of the public is certainly bad news.

1.5. DEFINITION OF HIGH-FREQUENCY TRADING

If, as will be seen later, the effects of HFT on market dynamics are unclear, defining HFT is not entirely clear either. According to the Securities and Exchange Commission (SEC), as reported in Friederich and Payne (2012), the very term High-Frequency Trading in not clearly defined, although it is being recognised to involve fast order submission, order modification (including cancellation) and extensive automation of the whole process. The same authors identify small positions, held for short periods in a proprietary capacity as features of HFT. Moreover, they highlight that HFT is not a trading strategy but rather a means of executing strategies, even though strategy crowding does not seem an unusual occurrence.

The SEC has dedicated quite some efforts to finding a suitable definition of HFT and identified in SEC (2014) at page 4 the following five characteristics often attributed to HFT:

(i) Use of extraordinary high speed and sophisticated programs for generating, routing, and executing orders

(ii) Use of co-location services and individual data feeds offered by exchanges and others to minimize network and other latencies

(iii) Very short time-frames for establishing and liquidating positions

(iv) Submission of numerous orders that are cancelled shortly after submission

(v) Ending the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions overnight).

1.5.1. IS SPEED THE ONLY DISCRIMINANT?

However, these characteristics describe how HFT behaves but do not define it. In fact, a trader would still be operating at high frequency if it makes use of a very high speed dedicated network rather than co-location, or if it uses to hold significant overnight positions and even if it only posts aggressive orders and is not involved in highly frequent submission and cancellation of limit orders. Whereas speed is a *conditio sine qua non* for HFT, all other features are accessory, perhaps very common but by no means necessary for a definition of the phenomenon. To make things even more complex, SEC (2014) reports that sometimes firms usually classified as non-HFT trade almost at the same speed as recognised HFT firms and do amend their orders as or more often. Not even co-location implies the HFT stamp on a firm, as the same document reports of numerous non-HF traders using co-location facilities. Nor other quantitative measures, whether order-to-trade ratios, number of fast messages, holding time, end-of-day position, aggressive vs. passive trading, seem to say the ultimate word about this matter. The conclusion of SEC (2014) is that HFT is not a monolithic phenomenon, but the sum of diverse trading strategies and behaviours. In this view HFT looks more like a possibility picked up when it suits the outstanding opportunity rather than a club whose rules are to be strictly adhered to, with expulsion to follow for any breach. A simpler definition of HFT is given by Barker and Pomeranets (2011): "High-frequency trading (HFT) is the high-speed execution of automated trading strategies, in which large numbers of trades are conducted in short time spans in order to profit from pricing or other market inefficiencies" (ibid. p.48), although this seems to exclude a rather common HFT strategy like small swing exploitation, which is not necessarily a market inefficiency. Philosophical disagreements have been identified by Zervoudakis et al. (2012) about a definition of HFT, seen as a technological evolution of traditional trading the same way as cars can be regarded as the successors of horse-drawn carriages. The paper also highlights that as physical limits are approached, innovation becomes commoditised and any further speed increase yields diminishing returns. Therefore, counter-intuitive as it may sound, market and sentiment analyses would reaffirm their predominance, irrespective of speed and technology, exactly as in the old open outcry days. There are a few authors who assimilate the CPU co-located at the exchange rack to the traders physically being on the open outcry trading floor where modern slower computer-based traders resemble the old off-floor players. A different opinion is held by Brogaard (2011), who recognises that those old days are gone forever and notices as, at the very minimum, since HFT introduction the size of orders and their resting time both decreased, whereas the number of orders increased. Yet, these are also characteristics of modern financial markets rather than a definition of what HFT is.

1.5.2. HIGH-FREQUENCY CANCELLING

Hasbrouck and Saar (2013) suggest that the expression 'high-frequency trading' for describing low-latency activity is generally a misnomer, in as much the practice usually referred to does involve high-frequency activity, like submitting orders, but it does not necessarily translate into trading. Indeed, the large majority of high-frequency limit orders get cancelled within a very short amount of time, and before they are executed. In some cases this is due to HF firms scouting the market in search for hidden liquidity. The jargon for this operation, 'pinging', comes from the computer communications utility called 'ping', which sends a short message to a remote computer to verify the status of its connection to the network. Analogously, a trader sends an order to the market to verify the existence of hidden quotes. If there are no concealed orders, the 'ping' order gets immediately removed. This strategy, together with a number of more or less manipulative practices (not addressed in this study), often creates the so-called phantom liquidity. Fast order submission and cancellation is a rather significant phenomenon as described by Zhang and Baden Powell (2011), according to whom three firms, accounting for 39.6% of orders submitted on the Paris CAC-40 list in April 2010 (and this is already an astonishing number) were found cancelling 96.5% of their orders. Similarly, Brogaard (2011) reports that according to 'some estimates', certain stocks experience up to 90 cancellations for every one trade executed. Quite interestingly for attempting to understand HF traders' behaviour, van Kervel (2015) reports that after a trade is executed, limit order cancellations on competing venues occurs rather often. This issue is dealt with by Angel (2014), who acknowledges traders' complaints about phantom liquidity and somehow compares it to the concept of Schrödinger's cat (from quantum physics): a market observers cannot tell whether or not a quote is 'alive' without interfering with the environment, in this case submitting a market order to spot hidden liquidity, that is, by changing the environment itself. Yet, not all scholars share this worry about phantom liquidity: Blocher et al. (2016) challenge its very existence, reporting that "little, if any, evidence on the question of phantom liquidity exists" (ibid. p.6). The phenomenon of order cancellation is currently being discussed thoroughly but no simple answer seems to exist. One of the most straightforward solutions proposed is to penalise high order-to-trade ratios (OTR). Friederich and Payne (2015) study the impact on market liquidity of the introduction of a penalty for high OTR implemented by the Italian Stock Exchange to curtail high-frequency quote submission. Their results do not support the view of the advocates of an OTR fee, leading to a collapse in the quote depth and an increase in price impact. The conclusion of the two scholars is that "the Italian OTR fee had the effect of making Italian stocks markets more shallow [sic] and less resilient" (ibid. p.214).

1.5.3. VOLUMES

HF traders do not accumulate large inventories and according to Kirilenko et al. (2011) their position in the E-mini S&P 500 futures during the first week of May 2010, oscillated between approximately +/- 3,000 contracts. As a matter of fact, a consequence of HFT entrance in the market noticed by Haldane (2011) is the rise in number, and fall in average size, of trades executed. Indeed, a parameter that is not easily noticed at 'normal' observation speed is volume. Zigrand, Cliff and Hendershott (2012) argue that even though daily volumes exchanged have risen since HFT entry, the second-by-second volume may not. This shapes market dynamics differently from pre-HFT times but again may not be taken as a discriminant between being a HF trader or otherwise. Lack of unique measurement criteria leads to wide variation in volumes measured. Even SEC (2010) recognises a somehow lack of clarity in the definition of HFT, something that makes the

Commission's task of reviewing the market structure issues more complicate like, for example, reliably estimating actual HFT volumes.

1.5.4. NOT ALL ALGORITHMIC TRADING IS HIGH-FREQUENCY TRADING

In another of its documents, SEC (2014), the US Securities and Exchange Commission notices the difficulty of separating Algorithmic Trading activities from proper HFT when using proxies like high message rates, bursts of orders cancellations and modifications, high order-to-trade ratios, small trade sizes, and increases in trading speed. The reason being that "HFT represents a large subset, but by no means all, of algorithmic and computer-assisted trading" (ibid. p.5). Similar concerns are raised by Friederich and Payne (2011) who highlight how "the majority of the early CBT [Computer-Based Trading] models traded agency flow in an aggressive fashion but without any particular high-frequency information advantage" (ibid. p.8). Among the many difficulties in properly defining HFT is that, aside of keeping their strategies carefully secret (as all trading firms do), HFT is not a continuous process but it rather tends to occur in micro-bursts, and their activities are kept confidential and anonymous by the exchanges.

1.5.5. CONCLUSION

As a conclusion from this review of the different approaches to reach a common definition of HFT, we can say that academics, practitioners and regulators could not find an unambiguous, commonly-agreed, clear-cut definition of what High-Frequency Trading is. Some trading practices are often adopted by HF traders but also by some non-HF traders - and conversely HF traders do sometimes fail to adhere to the 'standard' HFT practices. So, the question is still pending: What defines High-Frequency Trading?

Unfortunately, there seem not to be an ultimate word yet, and this ambiguity will be the unwelcome companion for the remainder of this study - as it is in academia’s and regulatory bodies’ discussions about HFT.

1.6. HIGH-FREQUENCY TRADING AND THE FINANCIAL MARKETS

1.6.1. SPEED

The classic trading theory assumes 'informed' investors. Letting aside, as is common knowledge, that there are 'noise' investors as well that are not informed, even among informed investors there are different degrees of 'informedness', as noticed by Darley and Outkin (2007): even before HFT time, some investors were constantly connected to a Bloomberg screen while others used to read regularly the Wall Street Journal. No one can claim the latter investors are not informed but the promptness of information is largely different between the two types of investor. This implies that speed is a major factor, if not the factor altogether, in today's trading environment.

A speech held by the Bank of England's Executive Director [Haldane (2011)] lists the three major effects HFT has on the market: (i) ever-larger volumes of trading are concentrated into ever-smaller time windows; (ii) traders take and execute their strategies at ever-higher frequencies; and (iii) market interactions tend to be ever more machine-to-machine rather than among humans. This fact is acknowledged by the CFTC-SEC (2010a) report, according to which "[m]any trading firms have strategies that are highly dependent upon speed in a number of areas: speed of market data delivery from exchange servers to the firm's servers; speed of processing of firms' trading engines; speed of access to exchange servers; and speed of order execution and response by exchanges" (ibid. Appendix B p.8). Speed-based strategies are indeed quite common: CFTC-SEC (2010a) estimates HFT volumes in the equity market to be often in the order of 50% or more of total trading volume. The same report evaluates the total number of daily trades originated by high speed systems (whether defined as HFT or otherwise) in over one million per firm. Given the high order-to-trade ratio of HFT, the number of trades is only a tiny fraction of overall messages. Hendershott (2011) reports NASDAQ receiving between 500 million and one billion orders per day and the entire US financial exchange system about five times more, with an average of 100,000+ events per second. Yet, most of those orders are never executed because they rest on the books a few milliseconds and then get cancelled. Not surprisingly, as noticed by SEC (2014), HF traders prefer competing on speed (as that is where their competitive advantage resides) rather than price. Moreover, according to Brogaard, Moyaert and Riordan (2014), HFT seems more profitable in periods of high volatility. This sounds intuitive as firms able to exploit tiny price oscillation ought to thrive in a volatile environment, although there is not unanimous consensus among academicians on this. On their side, Farmer and Skouras (2012a) identify those they regard as the *three and only* [italics in the text] sources of money in an order book market dominated by speed as: (i) picking off others' stale limit orders, (ii) preventing own limit orders from becoming stale, and (iii) beating competitors in obtaining the best position in the order book queue. It is easy to understand how all these sources of money are deeply influenced by execution speed. Brogaard (2011) to a certain extent agrees that, by exploiting their speed advantage, HF traders only have a limited number of avenues available for reaching profitability: market making, liquidity rebates, detecting statistical patterns and arbitraging. Acting as market makers is probably the most widely-adopted strategy by HF traders.

1.6.2. MARKET-MAKING

However, Foresight (2012) highlights some important differences between HFT and traditional market making: HF traders take advantage of their speed (often achieved via co-locating their computers at the exchange premises) to enter massive number of limit orders with the purpose of earning the bid-ask spread. Such orders are usually held for very short periods of time (from seconds down to microseconds) so minimising the risk of being picked off by more informed investors. In fact SEC (2014) recognises that HFT passive market making strategies may generate an enormous number of order cancellations or modifications - and certainly many more than for market taking strategies. Whereas traditional market makers carry affirmative obligations, HF traders do not. In particular, they have no obligation to provide quotes of a certain magnitude, with a maximum spread and for a certain amount of time over each trading day, as traditional market makers do. Yet, HF traders can make use of rebates granted by exchanges for providing liquidity. The importance of affirmative obligations is reported by Weaver (2012) as lowering spread and volatility and improving overall market quality. The obligation of providing two-way continuous quotations would also improve liquidity. But, how warned by CFTC-SEC (2011), increased competition among exchanges "has essentially eliminated rule-based market maker obligations" (ibid. p.2), potentially jeopardizing market stability. Lack of obligations is also worryingly noticed by several other sources, confirming it to be an issue for regulators, academics and practitioners alike. The main issue linked to lack of obligations is that HF traders may supply liquidity when it is cheap doing so, and walking away when it is dear and most needed by the market. This behaviour can, in stressful situations, harm liquidity and lead to instability.

1.6.3. MISCELLANEOUS IMPACTS

HFT has also been identified by many authors as the main driver of increased trading volume. Yet, Zhang and Baden Powell (2011) argue that "trading volume is not necessarily a reliable indicator of market liquidity especially in times of significant volatility" (ibid. p.10). The Foresight (2012) report acknowledges that although most academic studies show that HFT improves market quality, it is not clear if long-term investors have been positively or negatively affected. A generally positive role of HFT is recognised by CFTC-SEC (2011) as a factor for increasing competition, reducing transaction costs but it also warns about creating market fragility in highly volatile periods. However, even though Aitken et al. (2012) in general agree about the positive impact of HFT on market efficiency and integrity, they admit "our understanding of high-frequency trading and its implications on market quality are at best moderately informed" (ibid. p.3), where the expression "our understanding" seems to be more related to the entire academic community rather than solely to the authors. Indeed, opinions differ markedly. Danielsson and Zer (2012) accept the result that most HF traders follow price reversal strategies, therefore diminishing intra-day volatility, and according to Menkveld (2014) HFT benefits the market in the sense that information gets into prices quicker, although the same paper recognises that if "HFT uses [information] to aggressively pick-off other investors' quotes that are slow to change [...], it is likely to destroy welfare" (ibid.

p.16). If on one side Angel (2014) states that most HFT strategies tend to stabilise the market, on the other side more cautiously SEC (2010) warns that some HFT strategies may benefit market quality whereas other could be harmful. Technology advances have undoubtedly allowed shorter execution time and generally reduced trading costs but the increase in the level of electronic message traffic relative to trades has made life more difficult for those participants who cannot afford (or do not intend) to compete on speed. If on one side Gyurkó (2010) argues that fragmentation supports competition by narrowing spread and reducing transaction costs, on the other a significant positive correlation between variables that proxy HFT and market manipulation has been reported by Frino and Lepone (2012) - although Arnoldi (2016) also notices algorithms misled by human manipulation. Yet, it is also true, as noticed by Angel (2014), that the risk of market destabilisation and exacerbation of abnormal movements is a companion modern traders have to learn living with. Indeed, the non-linear sensitivity to change is a common worry as reported by Foresight (2012), which highlights how in the aftermath of the Flash Crash a 'butterfly effect' [Gleick (2008)] can be generated by the sole effect of speed, particularly when voluntary risk management actions, like stop loss, or regulatory enforced actions, like delta-hedging, may create directional upward or downward spirals with unpredictable outcomes. Because of this effect, markets may turn from liquid to illiquid in a matter of milliseconds, with all the negative consequences that made themselves evident on May 6, 2010. But unfortunately there are more unpredictable effects. Even a set of systems, otherwise stable on their own, may interact in highly unstable ways. The Report concludes that "the systemic risks of CBT are currently not very well understood" (ibid. p.30). Another kind of multi-disciplinary parallel is drawn with sociology, when the concept of 'normalisation of deviance' is borrowed to explain how any deviations from expected behaviour are regarded as temporary aberrations that will revert to the mean in due course. They are "seen as increasingly normal, until a disastrous failure occurs" (ibid. p.12). Also quite worried is the conclusion from Manahov, Hudson and Gebka (2014), who regard HFT strategies unrelated to the intrinsic value of the investment and only concerned in the asset's price in the next few seconds, strongly limiting any social value of their action. Another related and important characteristic of HFT is the arms race it gave rise to. Essendorfer, Diaz-Rainey and Falta (2015) recommend regulator, policy-makers and academics to analyse patent data (a sign of technological innovations) when trying to understand the roots of recent changes in capital markets. Technological innovation in the search of speed advantage leads to large investments for all players only to stand still at the same level of other competitors, just because 'the winner takes it all', with no social benefit arising from all that investment. According to Budish, Cramton and Shim (2015) the continuous-time serial processing way in which markets are designed implies that even public information create arbitrage rents. The attempts to exploit such opportunities leads to a never-ending arms race for speed which is socially wasteful. An example is provided by Laughlin, Aguirre and Grundfest (2014) who, using relativistically correct millisecond resolution tick data, find evidence of three millisecond reduction in one-way communication time between Chicago and New York (areas where some of the world largest exchange venues are located) during the period April 2010 and August 2012. Even the innovation HFT brings to the market is seen as myth rather than reality by many authors, among which Angel (2014) and Frino and Lepone (2012), in as much HF traders are often implementing rather old and fairly low-technology strategies, while their only focus is on brute speed. Not surprisingly HFT strategies are best-kept secrets but, since the technology is known and the effects of such strategies are evident it is not an impossible task to make an educated guess how HFT algorithms work. Ahlstedt and Villyson (2012) also agree that "strategies involved in HFT are often fairly simple" (ibid. p.3). If all the focus is on speed, then clearly the simpler the algorithm and the faster its execution. There are obviously several tricks of the trade that may help but they are known by all the players and once all of them have been adopted, all the latest technology purchased, and all the fine-tuning put in place, what remains is just to reduce the number of lines of code the algorithm goes through. And a lesser number of lines of code will fatally lead to simpler algorithms, with semantic diversification between one and the others gone at a large extent. An indirect confirmation of this is suggested by Chaboud et al. (2014), who find "potential for higher correlation in computers' trading actions than in those of humans" (ibid. p.1), whereas Farmer and Skouras (2013) report, though presenting no evidence, wide recognition by market participants of 'crowdedness of computer trading'.

1.6.4. FAIRNESS

Simple algorithms for faster execution not only provide an undeniable advantage versus humans but according to Cohen and Szpruch (2012) HFT is also handling well the market law of price risk through time, as its superior speed ensures the advantage. Moreover, the rather common practice of co-location risks to jeopardize fairness to other players. Indeed, as pointed out by Angel (2014), proximity to the exchange matching engine allows some participants to receive market data before others do - and therefore to submit orders in advance to the slower traders. In many cases the latter traders do not even have the chance to notice an order because it is received by the exchange, acted upon and executed before they can even blink. This practice is prone to create situations never seen before in many centuries of market activity. The speed at which orders are entered and cancelled, and the extremely short interval between these two complementary events related to the same instrument, depict a market in which perception of depth is illusory. The reasons are: (i) an observed quote may suddenly disappear before an investor is able to place a market order for it, and (ii) "if several market orders hit a financial asset that appears to have many limit orders near the inside price, there is no guarantee that the apparent market depth will still be available on the order book after the first trade executes but before the second one arrives" [Brogaard (2011), p.5]. Moreover, the same paper also highlights another issue, an old problem potentially made more acute by computer-assisted orders. When a market order gets executed, it consumes liquidity and it sometimes clears the price level at which it traded altogether. At 'human' speed this occurrence may occasionally be noticed as barely annoying, but at near-light speed sub-optimal price trading may become an unwelcome, yet constant, companion of lower-technology investors. "The concern now is that limit orders change so fast that for most markets participants the quoted price at the time of making a trade decision is not indicative of the price at which the trade will take place" (ibid. p.9). A fascinating hypothesis of the reason for HFT market environment being so radically different from the one we were all used to, is suggested by Hasbrouck and Saar (2013), drawing a comparison to "the challenge faced by physicists when attempting to relate quantum mechanics' subatomic interactions to our daily life that appears to be governed by Newtonian mechanics" (ibid. p.21). Because of this change of paradigm - and following the widespread concern due to the Flash Crash, it comes as no surprise that HFT has been put under scrutiny by regulatory authorities both in the US and in Europe.

1.7. STRUCTURE OF THIS RESEARCH

As its title suggests, this research focuses on two main topics, High-Frequency Trading and financial stability. Therefore, these two concepts, and their relationship, will be the parameters against which every aspect dealt with herein shall refer to. Many studies in the literature highlight the positive impact of High-Frequency Trading on the main market parameters relevant to financial stability: Conrad, Wahal and Xiang (2015) find that High-Frequency activity lowers transaction costs; Brogaard, Hendershott and Riordan (2014) show that its contribution to price discovery is statistically significant; Myers and Gerig (2014) notice that it reduces volatility, whereas Chaboud et al. (2014) also find a beneficial effect on liquidity. Yet, other authors, for example Aldridge and Krawciw (2015) find a correlation between aggressive High-Frequency activity and volatility and Kirilenko et al (2011) recognise a different impact according to market conditions. The difference of opinions suggested further research on the topic, in particular the use of computer-based simulations for assessing the potential impact of High-Frequency Trading on some financial stability parameters and audit trail data analysis for verifying whether the results of the simulations are compatible with the data.

Each of the following sections will describe the content of the corresponding chapter in the thesis.

1.7.1. LITERATURE REVIEW

Chapter 2 reviews the literature. It starts by describing the financial stability parameters identified earlier in this chapter, as the ones most relevant to trading, in their relation with HFT. Since the current literature shows very different results in nearly all aspects of HFT investigated, there seem to be a lot of room for research. In particular some articles praise the role of HFT in mitigating market volatility whereas other papers reckon an increase in volatility since HFT entered the scene. Moreover, some results are different according to whether markets are quiet or in turmoil. Then the following paragraph (2.2) reviews the literature on how those parameters were affected during, and on their turn affected, the most dramatic event allegedly ascribed to HFT: the May 6, 2010 Flash Crash. That event sparkled a large amount of hypotheses about what caused it and many of them are in conflict with others. There seem not to be much agreement on this topic either, therefore further research seems appropriate. Many of the hypotheses concerning the Flash Crash are presented in paragraph 2.2, together with a review of the all-important topic of market microstructure. Paragraph '2.3. Literature Review on Market Tiering' addresses the literature about a little investigated aspect of financial instability, the possibility that HFT leads to the market splitting into two tiers, one reserved for fast participants and the other for snail-trading, implicitly riskier. It is not completely clear if HFT actually tiers the market and even those papers suggesting it does, fail to provide quantitative evidence (a gap addressed by this research). This topic has relevance to the concept of risk in general and to the fairness for all participants, both aspects being related to the wider issue of financial stability. Then the chapter continues with paragraph 2.4 on the impact of HFT on arbitrage and on its relationship with the Efficient Market Hypothesis, another topic vastly debated in the literature although lacking quantitative evidence (another gap addressed by this research), before concluding with the identification of the gaps in the literature (paragraph 2.5), which directly links to the description of the research in the following chapter.

1.7.2. THE RESEARCH

Chapter '3. The Research' links to the gaps in the knowledge so far to introduce the research (section 3.1.1), describes its purpose (section 3.1.2) and how this thesis is going to addresses the five areas of analysis (sections '3.1.3. Impact of HFT on Price Volatility', '3.1.4. HFT and Market Tiering', '3.1.5. Impact of HFT on the Efficient Market Hypothesis', '3.1.6. Data Analysis of the Flash Crash', and '3.1.7. Relative Versus Absolute Speed') with the goal of better understanding the HFT phenomenon and its impact on financial stability. The last four paragraphs of chapter 3 explain the reasons for the main choices made in this thesis: paragraph 3.2 explains the benefits of using simulation, the technique widely adopted in this research, as identified by the literature on the philosophy of science; paragraph 3.3 explains the reasons for selecting, among the many possible, the Chicago Mercantile Exchange as the environment for the simulations and the analysis; then the reasons for adopting a sequential simulation versus a parallel one (paragraph 3.4) and for choosing Petri Nets as a modelling tool (paragraph 3.5) are also described.

1.7.3. IMPACT OF HIGH-FREQUENCY TRADING ON VOLATILITY

Chapter '4. Impact of High-Frequency Trading on Volatility' addresses one of the most contrasted topics identified in the literature. The chapter describes in details the first simulation, used to test the impact of HFT on market volatility. It presents its methodology (paragraph 4.2), together with some simplifying assumptions, then it describes the details of the simulation (in paragraph 4.3) and finally presents and discusses the results (paragraph 4.4). This chapter adds a contribution to the topic, which the current literature fails to settle in a definitive way.

1.7.4. DOES HIGH-FREQUENCY TRADING CREATE A TWO-TIER MARKET?

The following chapter focuses on the possibility of market tiering in presence of HFT, a topic little investigated in the literature and even when it is, lacking quantitative measurements. The structure of the chapter is similar to the one adopted in chapter 4. Before presenting the methodology (paragraph 5.3) and describing the simulation (in paragraph 5.4), paragraph '5.2. Is Market Tiering a Threat to Financial Stability?' analyses the relationship between endogenous risk and phase transition, during which allegedly the market abruptly modifies its behaviour, and why and how this may split the market in two tiers. Section '5.5.1. Main Approach' presents the results achieved by simulating both the trading strategies usually preferred by fast traders and their speed difference compared to slow traders. The following section ('5.5.2. Alternative Approach') releases the speed difference feature as a counter-proof for assessing whether speed had a role. Obviously similar results with both approaches would cast doubt on speed difference being the main cause of market tiering. Paragraph '5.6. Discussion' highlights the meaning of the results together with some concerns they raise and paragraph '5.7. Conclusion' relates the results to the concept of financial stability.

1.7.5. HIGH-FREQUENCY TRADING AND THE EFFICIENT MARKET HYPOTHESIS

The last simulation, presented in chapter 6, addresses another important aspect of the potential impact HFT has on the classical economic theory, namely the Efficient Market Hypothesis ('the Hypothesis'), and consequently the impact arbitrage in a HFT environment has on financial stability. To this end, the interpretation of the Hypothesis is assessed against the technological changes occurred over the last decade or so (sections '6.2.2. Arbitrage and Transaction Costs' and '6.2.3. Is Arbitrage a Real Issue?'). Several articles state that HFT may actually affect the validity of the Hypothesis but little quantitative findings are displayed in the current literature. Then, the current HFT practices in relation to arbitrage are investigated in detail ('6.2.4. Do Arbitrage Opportunities Exist in the Real World?') before presenting another simulation (in paragraph 6.3) to verify whether the ground upon which financial knowledge has been built in the last fifty years is still solid or it may find itself under question in the new HFT-dominated environment. The chapter concludes by challenging its own approach, discussing its results and evaluating their impact.

1.7.6. FLASH CRASH DATA ANALYSIS

The data analysis carried out in chapter '7. Flash Crash Data Analysis' focuses on the extreme event that in May 2010 spread panic throughout the financial world. The goal of the chapter is spotting out the main causes of the Crash and providing a link between some theoretical results found in previous chapters and hard data. It does so by analysing in detail the audit trail data from the critical day and the days nearby, in order to highlight, by adopting an original viewpoint (analysis of the 'runs') and with the help of different statistical methods, the most significant differences that led to the crisis. In particular, the analysis links the findings of chapter 4 with real data to provide evidence that the so-called naïve orders led to excessive volatility. It also links Stop Loss orders to the little explored concept of absolute speed to illustrate the issue of frequent order cancellation, widely discussed in the literature but (although with some notable exceptions) not always by providing evidence of an exacerbating factor of the crisis.

All results found in the four chapters presented above and in the literature are summarised, compared, confronted and discussed in a comprehensive way in chapter '8. Conclusion'. The last paragraph (8.3) proposes possible refinements and how future research on the above topics could proceed.

1.8. CONCLUSION

It looks like there is not much agreement among experts on what High-Frequency Trading is. The concept is intuitively easy to grasp but when a precise definition is called for, problems arise. As far as HFT is concerned, from a financial stability point of view, it does not really matter whether a firm is deemed to belong to that category, it is the operational behaviour at any instant in time that makes the difference. In fact, this will be the approach pursued throughout this research. Only slightly more agreement seems to be reached on what financial stability is and Alawode and Al Sadek (2008) notice, as a sign of the increased attention dedicated to the topic of financial stability, that several international forums devoted to this issues have emerged or have recently become more active. The authors mention the Financial Stability Forum, Basel Committee on Banking Supervision, Financial Stability Institute, Committee on the Global Financial System, Committee on Payment and Settlement Systems, International Association of Insurance Supervisors, International Accounting Standards Board, International Organization of Securities Commissions and the

International Association of Deposit Insurers. Moreover, the Counterparty Risk Management Policy Group is a private sector organization also devoted to fostering financial stability. The very same fact that so many organisations are needed to try and ensure achievement of financial stability raises doubt the goal is still hard to reach. However, as Danielsson (2013) sensibly notices, these committees may actually have done an excellent job: we don't see the successes, we only see failures. A common place after each financial crisis is the 'this time is different' mantra. It may well be the case, but "[f]ighting crises is like fighting bacteria. We are able to develop a medicine that fights the parasite in the short run, but eventually the bacteria evolve and the medicine becomes ineffective. It is the same with policy: we can prevent the old crises from recurring, but the next one will simply take a new form" (ibid. p.80). That's why keep fighting for financial stability is so important.

In this chapter the surface of HFT has been scrapped. The definition and its relation with the markets tells little about the real, or presumed, impact it has on the market parameters identified as most relevant to trading. This is where the next chapter starts from.

2. LITERATURE REVIEW

This chapters presents the four main streams of literature related to the four streams of this research. The next paragraph illustrates the features of High-Frequency Trading especially in relation to the financial stability parameters. Paragraph 2.2 describes the May 6, 2010 Flash Crash and summarises some authoritative opinions on its possible causes. Moreover, some studies suggest that ultra-fast trading might lead to HF traders dealing mostly with each other in case of narrow spread. This little investigated phenomenon could lead to market tiering and the relevant literature is reviewed in paragraph 2.3. The following paragraph (2.4) deals with the topics of arbitrage opportunities and the impact of HFT on the Efficient Market Hypothesis. The last paragraph concludes.

2.1. LITERATURE REVIEW ON HIGH-FREQUENCY TRADING

2.1.1. INTRODUCTION

Despite being a relatively recent phenomenon, HFT has attracted a copious amount of literature, especially in the aftermath of the Flash Crash on May 6, 2010. The main purpose of the academic studies on HFT is to set a scientific framework in order to understand whether or not HFT is harmful to the orderly behaviour of the financial markets, and if so, to what extent and, above all, why and how. The effect of HFT on financial markets is controversial according to Farmer and Skouras (2013), who base their statement on several academic papers and studies published between 2009 and 2011 that reach very different conclusions. A similar view is shared by Sornette and von der Becke (2011), who refer to studies finding a positive correlation between HFT and beneficial market effects, like increased liquidity, reduction of transaction costs, lower spread and volatility whereas, on the other side, they also mention other studies supporting the opposite view about the role of HFT being a destabilising factor.

The next sections set the framework by reviewing the literature about the impact of HFT on the main trading parameters: transaction costs, price discovery, volatility, bid-ask spread, liquidity, market making, HFT profitability.

2.1.2. TRANSACTION COSTS

Many researchers do claim findings about HFT – and the majority of them tend to regard HFT in a rather favourable light. Foucault, Kadan and Kandel (2013) develop a theoretical model describing the interactions between market makers and market takers, and find that algorithmic trading directly increases trading rate via reduction in monitoring costs. Many studies support the view that HFT is not harmful to the stability of financial markets and even to the traditional investors. By focusing on the effective spread costs in the Ancerno dataset, Brogaard et al. (2014) find no measurable effect by the increase in HFT activity on execution costs for institutional investors. More boldly Kirilenko and Lo (2013) use stochastic dynamic programming evaluating the expected cost-minimising sequence of trades to deny any doubt that algorithmic trading yields "tremendous cost savings, operating efficiency and scalability in every financial market it touches" (ibid. p.52). Conrad, Wahal and Xiang (2015) analyse 2009 Trade And Quote (TAQ) data stamped at second granularity, and 2010-2011 TAQ National Best Bid Offer data stamped at millisecond, finding that, on average, high-frequency activity significantly lowers trading costs. More cautious but still pointing into the same direction is Menkveld (2013). Investigating the impact on trading fees by new high-tech entrant markets specifically suited for HFT, like BATS in the US and Chi-X in Europe, the study finds large cost reduction, even though competition among markets to secure the larger share of trading seems to play a crucial role in this case. Also positive about trading cost reduction, albeit rather concerned about collateral damages brought by HFT, is Harris (2016), who in particular focuses attention on systemic risk, front running, and reduced investor confidence. More negative about the impact of HFT on transaction costs are Ding, Hanna and Hendershott (2014). The authors find several price dislocations between different NBBO data feeds every second and such dislocations last no more than two milliseconds, causing extra costs for slow traders frequently active on the market, like institutional investors. A similar opinion is expressed by Hoffmann (2014), who notices how slow traders submit limit orders which have lower execution probability because of the presence of fast traders, and "[b]ecause speed is a source of market power, it enables fast traders to extract rents from other market participants" (ibid. p.156).

2.1.3. PRICE DISCOVERY

The Brogaard et al. (2014) research mentioned above also states that HFT shows the positive effect of increasing price efficiency as HF participants tend to trade in the direction of permanent price change, so speeding up price discovery, and in the opposite direction of temporary mispricing, both on days showing average and above average volatility. Manahov and Hudson (2014) confirm efficiency in the price discovery process using one-minute high-frequency data from the six most often traded currency pairs (USD/EUR, USD/JPY, USD/GBP, USD/AUD, USD/CHF and USD/CAD). Similar conclusions are also reported by Hendershott (2011): the referred study analyses price efficiency trends over a period of four years (2006-2009) at NYSE and NASDAQ and argues that, based on intraday variance ratios, growing HFT activity leads to higher market efficiency (which includes quicker price discovery). Frequent price cancellations are also a valuable source of information for price discovery purposes, according to Blocher et al. (2016), who base their research on 5.78 terabytes of data on all the S&P stocks for 2012. "The HFTs process the information so quickly that price discovery comes from the cancelations rather than from executions. This is the more effective method, since no dollars need change hands" (ibid. p.8). Aitken et al. (2012) study cross-sectional determinants of HFT participation over long time series on the Euronext Paris and London Stock Exchange (LSE), finding that “the increase in the level of HFT activity has increased efficiency without harming the integrity of the market” (ibid. p.40). Using data provided by NASDAQ under non-disclosure agreement and by NYSE, Brogaard, Hendershott and Riordan (2014) show that HFT contribution to price discovery is statistically significant and that HFT is negatively correlated to pricing error. Pricing errors, noise and volatility, together with narrow spreads, are identified by Benos and Sagade (2016), who analyse UK markets' intraday behaviour of HFT, as a favourable environment for HFT to trade relatively more. Aitken, Cumming, and Zhan (2015), studying a sample of 22 stock exchanges in 18 countries ranging from Malaysia and China to the US, find a significant mitigation in the frequency and severity of end-of-day price dislocation due to the presence of HFT. Buchanan (2015) praises HFT for its contribution in synchronising prices across markets. Rather critical of HFT's merits seem to be Stiglitz (2014), who fails to see any social usefulness in HFT, and Paul Krugman, mentioned in Linton and O'Hara (2012), argues it being "hard to imagine a better illustration (of social uselessness) than high-frequency trading. The stock market is supposed to allocate capital to its most productive uses, for example by helping companies with good ideas raise money. But it's hard to see how traders who place their orders one thirtieth of a second faster than anyone else do anything to improve that social function" (ibid. p.29). On the same side sit Hasbrouck and Saar (2013), who cast doubt over the societal benefits of low-latency trading. Eliminating transient price disturbances certainly has its value "but such an argument at the millisecond environment is a bit tenuous" (ibid. p.15). Less straight but substantially aligned on this opinion are Benos and Sagade (2016), who state a direct relationship between price discovery and volatility occurring as a response to information about fundamentals, and an inverse one with excess volatility. In contrast with other researchers, Jung Lee (2015) bluntly declares HFT "detrimental to the price discovery process" (ibid. p.31). Overall, albeit with some exceptions, it looks like academicians' opinions converge on the fact that HFT does accelerate price discovery but diverge markedly on the benefits of such promptness. Moreover, if benefits exist at all, they seem rather unidirectional towards HF traders and against the rest of market participants. This point shall be further developed when discussing the impact of HFT on market efficiency (chapter 6).

2.1.4. VOLATILITY

By modelling a continuous double auction, a widely used structure shown by nearly all exchanges, an agent-based simulation produced by Myers and Gerig (2014) shows a drop in volatility as well as an indication that HFT provides liquidity to the market and Hasbrouck and Saar (2013) observe that more low-latency activity implies lower short-term volatility. The test of algorithmic agents' impact on volatility in a modelled market made up of informed, momentum and noise traders, leads Gsell (2008) to conclude that lower latency yields a statistically significant reduction in volatility. Analysis of the thirty most traded stocks on the Swedish index OMXS30 on the NASDAQ Stockholm exchange suggests Hagströmer and Nordén (2013) that HFT mitigate intraday price volatility. This result is consistent in both highly volatile months (data from August 2011) and much less volatile ones (February 2012). As a counterproof, the same research investigates the opposite hypothesis, verifying that "a decrease in trading activities of the opportunistic HFTs causes an increase in stock return volatility" (ibid. p.36). An analysis based on electronic trade-by-trade data provided by the InterContinental Exchange, Eurex, NYSE Euronext, and the CME Group by Bollen and Whaley (2015) yields similar results for the futures market. Kalejian and Mukerji (2016) use a sample of daily end-of-day prices from the 200 most traded S&P 500 stocks for the period 1/5/1985 through 31/3/2012, finding that "HFT has had a significant effect on daily asset price volatilities […and] the relationship between stock fundamentals and volatility has weakened since the advent of HFT" (ibid. p.88). Rather more cautiously Zervoudakis et al. (2012) analyse whether high frequency quoting does reduce bid-ask spreads and lowers volatility, finding "no direct and unambiguous evidence of causality between HFT and increased volatility" (ibid. p.7). They also argue that "[i]f HFT contributed to volatility, then HFT diffusion should have increased the intraday-to-overnight volatility ratio; but this correlation is not evident" (ibid. p.7) and Brogaard (2010) only finds "not strong" results of HFT having impact on reducing volatility or none altogether. The same weak or lack of evidence is empirically found by Groth (2012) about HFT withdrawal during periods of high volatility on the Xetra platform at the Frankfurt Stock Exchange. Zigrand, Cliff and Hendershott (2012) also fail to find direct evidence of a positive impact of HFT on volatility but they argue that, under some circumstances, CBT can trigger self-reinforcing feedback loops. This topic is further developed by Abrol, Chesir and Mehta (2016), who admit that positive feedback loops can amplify shocks and pose systemic risk, as events occur well beyond human reaction time - and even beyond human comprehension. Definitive results in favour of increased volatility, and in general of a dysfunctional role of HFT, as the number of HF traders rises is found by Jarrow and Protter (2012), who investigate the exploitation of arbitrage opportunities by HF traders. Similar trading strategies may lead to crowding effect and therefore to exacerbate price movements. This possibility is also investigated by Brogaard (2010), who interprets the results found as a confirmation that HFT engage in less diverse strategies than non-HFT, implicitly recognising HFT activity as a potential volatility factor. Indeed, a significant correlation between HFT strategies is also confirmed by Chaboud et al. (2014). Even according to Kirilenko et al. (2011), HF traders seem to have common strategies; they are found to follow the trend in the four seconds after a price showing a clear direction and going contrarian after ten seconds. This seems another suggestion of the crowding effect identified above which may increase volatility. Zhang [2010] uses exogenous shocks of NYSE autoquote to HFT and after taking into account firm's fundamentals and other exogenous volatility drivers, finds a positive correlation between HFT and volatility, especially for large capitalisation stocks. Aldridge and Krawciw [2015] find a correlation between aggressive HFT activity and volatility, although there is uncertainty about the causal relationship: "[i]t is not immediately clear [...] whether aggressive HFTs seek out high volatility, whether aggressive HFT participation induces higher volatility in stocks, or both". Once again, the academic world seems to reach diverse views on HFT impact on one aspect of market stability.

2.1.5. BID-ASK SPREAD

Another focus of market stability is the bid-ask spread, that is, the cost traders incur in buying at the ask price and selling at bid price. The impact of HFT on market quality is deeply investigated by Brogaard (2010), who tests whether HF traders flee in volatile markets (so failing to provide liquidity when it is required most) by analysing their activity as volatility increases and during varying degrees of 15-minute period price changes. The paper reports that HF traders often book the best bid and ask, especially for larger firm size. Jarrow and Protter (2012) adopt a theoretical model by using two equations, one representing stock price process in absence of HFT and the other in its presence. The research somehow cautiously states 'preliminary evidence' suggesting that HFT narrows spread, improves liquidity, decreases volatility and makes markets more efficient, even if it affirms no final verdict on this matter. A similar opinion shows Menkveld (2013), according to whom stocks affected by HFT activity experienced spread reduction by 30% in a year with respect to other stocks. A clear bid-ask spread downward trend and a contemporaneous depth upward trend is the result found by Friedrich and Payne (2011), who plot LSE best spreads and book depth for the FTSE 100 stocks between January 2009 and April 2011. The conclusion is that the authors “do not see which forces other than the growth of CBT could explain these trends” (ibid. p.20). The gap between CBT and HFT is filled by stating that “if CBT has contributed something genuinely new to markets, it is arguably a type of participant loosely called the ‘HFT shop’” (ibid. p.7). The reduction of the spread in ten emerging markets has been investigated by Yilmaz et al. (2015); using daily data for 361 stocks, panel data regression show that technological innovation, among which HFT plays a major role, decreased the bid-ask spread. A critical aspect is investigated by Menkveld and Zoican (2016), who develop a model with some HF traders providing liquidity and some consuming it. They find that spreads fall when limit orders execute at higher speed than market orders. The lowering spreads result is described and discussed, by providing more logical reasoning than hard evidence, by Gowan (2010), who also lists other benefits brought in by HFT: added liquidity, faster execution and cost reduction, though mainly brought by competition among HF traders than by HFT as such. Formal statistical hypotheses are instead tested by Menkveld and Zoican (2016), in particular whether latency reduction leads to increase of the adverse selection component of the bid-ask spread. The paper regresses the adverse selection component of the spread aggregated across stocks and finds that a drop in market latency has a positive significant effect of about 7%. However, most researches tend to align with the viewpoint of HFT increasing market stability, via spread reduction as well as increasing liquidity and dampening volatility. Zervoudakis et al. (2012) follow the main stream of thought by arguing that “HFT systems reduce […] spreads by allocating liquidity” (ibid. p.6). By studying cross-sectional determinants of HFT participation in LSE and Euronext Paris between 2001 and 2011, Aitken et al. (2012) find that a higher level of HFT activity tends to reduce bid-ask spreads. Ordinary NASDAQ book order data for Q4 2007 and June 2008 is used by Hasbrouck and Saar (2013) to estimate regression coefficients suggesting that “higher low-latency activity implies lower posted and effective spreads, greater depth, and lower short-term volatility” (ibid. p.25). A different result is reached by Hendershott and Moulton (2011) who analyse data from NYSE and the CBOE in the year around hybrid activation running from June 1st, 2006 through May 31st, 2007. The research finds that the change, which reduced execution time from about 10 seconds to less than one, increased the bid-ask spread because of the increase in adverse selection.

Although not unanimous, academic research shows a clear direction about a positive HFT impact on bid-ask spreads.

2.1.6. LIQUIDITY

In the aftermath of May 6, 2010 HFT became a target for charges of liquidity consumption, on its turn leading, under severe stress conditions, to stub quotes being displayed. The Myers and Gerig (2014) simulation builds a continuous double auction agent-based model in presence of HFT and finds better liquidity with no higher volatility. It also reports a higher probability of transactions to happen, which the study interpreted as another sign of higher liquidity directly caused by HFT. Benos and Sagade (2016) analyse transaction data about four stocks picked up from the FTSE 100 dataset at the one-second granularity and finds that overall HF traders tend to supply somewhat more liquidity than they consume. Moreover, they supply such liquidity when it is expensive doing so, and consume it when it is cheap [Hendershott and Riordan (2013)]. On the same line of thought, Hagströmer and Nordén 2013) subdivide the traders into HFT and non-HFT, and their model shows the HFT group supplying more liquidity than they consume and HF participants trading relatively more when markets are volatile. Yet, according to the paper, although these results are statistically significant, no evidence can be drawn relative to a cause-effect relationship: it is just stated as a fact. A sharper standing is taken by Carrion (2013): based on a NASDAQ sample of 120 stocks, the paper recognises that "HFTs provide liquidity when it is scarce and consume liquidity when it is plentiful" (ibid. p.710). Chaboud et al. (2014) analyse the impact of algorithmic trading in the ForEx environment and find that it is beneficial for both liquidity and volatility. A similar result is reached by Brogaard, Hendershott and Riordan (2014), who point out how HF traders supply liquidity in both high and low volatility days. Even stronger are the results found by Jarnecic and Snape (2014), who find HF traders to resolve temporal liquidity imbalances, and Brogaard, Moyaert and Riordan (2014), who report 18% higher HFT activity during jump interval compared to other periods, whereas Brogaard (2010) observes that in times of abnormally high volatility HFT do not stop providing liquidity, either on the bid or on the ask side. The paper also acknowledges that “HFT activity increases with a shock to volatility” (ibid. p.30). Focusing on the relationship between market liquidity of futures traded on EUREX Exchange and HFT activity on the European derivative markets, Hruska and Linnertova (2015) find a positive effect of HFT on market liquidity, despite somehow mixed results due to how volatility is measured. At the contrary, although Diaz-Rainey, Ibikunle and Mention (2015) acknowledge the role of HFT (and Exchange Traded Funds) in making trading in capital markets cheaper, they view liquidity as fragmented and opaque. Baron, Brogaard and Kirilenko (2012), using transaction-level data for the E-mini S&P 500 futures for the period August 2010 through August 2012, find HFT earning substantially higher profits when consuming liquidity than when providing it. This seems a strong reason for HF traders to diminish the available liquidity rather than increasing it. Somehow of the same opinion seem to be Cvitanić and Kirilenko (2010), who develop a mathematical model showing that during a crisis HF traders provide less liquidity than in normal times. Again, the judgement of Jung Lee (2015) about the role of HFT in the Korean futures market between April 2009 and March 2010 is unusually sharp: "high frequency traders (HFTs) do not provide liquidity in the futures market, nor does HFT have any role in enhancing market quality" (ibid. p.31). In the middle of the two parties, Buchanan (2015) states pros and cons. If HFT makes it easier for investors to find counterparts ready to trade at a mutually agreed price, HFT-supplied liquidity is fleeting and may become unreliable when markets get unruly. Substantially on the same cautious position are Goldstein, Kumar and Graves (2014) who contrast the early positive evidence of HFT as liquidity provider, with the more recent concern about fairness, like tools or rights given to HF traders that are unavailable to other types of investors. In the literature there seem to be some consensus, although not unanimous, on HFT being beneficial for liquidity supply. This fact is summarised by Aitken et al. (2012) who observe that "[d]espite these largely theoretical negative market quality findings, the majority of papers that deal with HFT empirically find a predominantly positive overall impact" (ibid. p.8). Barker and Pomeranets (2011) summarise a common view as "HFT appears to be having a profound impact on market liquidity, and its rise has coincided with an increase in trading volumes, tighter bid-ask spreads and lower market volatility" (ibid. p.48).

2.1.7. MARKET MAKING VS. TAKING

Many sources have indicated HFT as either the originator or the main culprit of the May 6th Flash Crash and in particular of the dramatic liquidity scarcity experienced on that day. However, an in-depth study carried out by Menkveld (2013), based on trade and quote data on Dutch local index stocks for both Chi-X and Euronext in the period 1/1/2007 through 17/6/2008, finds that "in both markets the vast majority of HFT trades are passive: 78.1% in Euronext and 78.0% in Chi-X" (ibid. p.22). A compatible result is found by Hagströmer and Nordén (2013), who compute order-to-trade ratio (number of limit order divided by the number of executions) leading to the conclusion of HFT acting most often as market makers. Moreover, they observe 63-72% of HFT trading volume being made of market making and 81-86% of limit order traffic, compared to arbitrage and momentum strategies. The discrepancy between the two ranges is due to the presence of fleeting limit order, whose features make them alike to market taking strategies. Hasbrouck and Saar (2009) call “fleeting limit orders” those limit orders cancelled within two seconds of their submission. They are close substitutes for market orders in as much their purpose is seeking immediate execution without running the risk of being picked off by a fast moving adverse market trend. Since some limit orders may not be publicly visible, one of the goals of fleeting orders is to pick up hidden liquidity. The same study reports 83% of all incoming orders in the Island ECN being limit orders but only 18.4% getting fully or partially executed. Many limit orders are cancelled within an extremely short time and 27.7% are cancelled within 2 seconds of submission. There is no sign of passive patience in trading strategies like this. HFT has changed the scenario completely. Kirilenko and Lo (2013) translate these figures into hard money by arguing that trades who enjoy best access to customer order flow (and best algorithms) earn top rewards. Trading is a business in which information means money: Benos and Sagade (2016) group HF traders according to market making vs. taking activity and their result "suggests that HFTs who pursue strategies that require the use of aggressive trades are most informed" (ibid. p.1). Another advantage of HFT market making is underlined by van Kervel (2015) who argues that “high-frequency traders who operate as market makers can strongly benefit from the increased execution probability. That is, placing duplicate limit orders increases their trading rate and expected profits” (ibid. p.7). Obviously, a slow trader could not afford the luxury of placing multiple limit orders in the hope that only one gets through execution as the risk of going off balance with multiple (unintended) executions before being able to cancel the other limit orders would be far too high. Jovanovic and Menkveld (2016) assume that HF traders are both faster and more informed than their counterparts. Based on this assumption their findings on market efficiency are mixed. Quite boldly Kirilenko and Lo (2013) state “[i]n contrast to a number of public claims, high-frequency traders do not as a rule engage in the provision of liquidity” (ibid. p.60). The same concept appears in a slightly different manner in Baron, Brogaard and Kirilenko (2012), where the authors notice how liquidity-taking HFT are especially profitable and their aggressiveness consistent across days. Barrales (2012) somehow rhetorically asks: "Are equity markets vulnerable to a sudden collapse if the traders who account for about half of the volume have no regulatory obligations to stabilize prices?" (ibid. p.1195). Once again, it seems that serious academic research leads to very different results, adding to the impression of understanding HFT being a tough nut to crack.

2.1.8. PROFIT OF HF TRADERS

A critical issue about HFT is whether this practice on its own ensures a profit or otherwise. If HF traders make money then it is plausible to investigate the impact of the ultra-fast trading on the regular functioning of the markets and the nuances of the high-speed algorithmic strategies. If not, that is, if HF traders do sometimes make and at some other times lose money, the conclusion would be to consider them as ordinary traders, subject to the usual random walk of the market. Once again, academics display a wide range of opinions. Buchanan (2015) notices how "profits earned by highfrequency firms have fallen in recent years" (ibid. p.163) and Serbera and Paumard (2016) point out the increased competition among high-speed trading algorithms and even "professional human traders adapting and building adequate responses", all leading to "shrinking profits for HFT" (ibid. p.271). The dataset analysed by Menkveld (2013) shows that all HF traders’ earnings arise from passive orders and that in general they lose money on their positions but make money on the bid-ask spread and this looks consistent across the stock universe analysed. Moreover, the profit component of the inventory seems to be restricted to trades closed within 5 seconds whereas virtually all those lasting longer than one-minute yield negative results and those in the middle are mixed. The rationale behind this seems to be the adverse selection of participant posting limit orders. A similar result is reached via the analysis of NASDAQ and NYSE in 2008-2009 by Brogaard, Moyaert and Riordan (2014), according to which HF traders lose on limit orders and gain on market orders. The paper finds that, at 90% confidence interval, HF traders incur losses during price jumps while traditional traders seem to profit from that. A question is natural at this point: why should HF traders engage in liquidity supplying activities if they lose money? The answer resides in the market fee structure. Foucault, Kadan and Kandel (2013) report trading fees for five US trading platforms (NYSE Arca, NASDAQ, BATS, EDGX, LavaFlow) ranging from -$0.30 to -$0.20 for make fees (that is, liquidity suppliers earn a few cents) per 100 shares traded, and +$0.25 to +$0.30 for take fees. This supports the view of HF trsders losing money on limit orders as they get compensated by the incentive. However, free market is as various as you can think and indeed Foucault (2012) reports make and take fees for a set of 10 US markets, showing that some exchanges offer no rebate for limit orders and there is even one exchange that offers -$0.18 take fee (rebate) against a +$0.14 make fee, both per 100 shares.

2.1.9. CONCLUSION

If one conclusion can be drawn from the literature on HFT, it is that opinions are rather diverse among academicians, and the final word seems far from being said. As noticed by Harris (2016): "The debate about HFT has been quite emotional, in large part because people naturally fear what they do not understand well" (ibid. p.6). The foreword of the Foresight (2012) final report highlights the fact that “some of the commonly held negative perceptions surrounding HFT are not supported by the available evidence” (ibid. p.5). Indeed, the same report, referring to the existing literature, goes even further by stating that “[r]esearch so far provides no direct evidence that HFT has increased volatility, not least because it is not clear whether HFT increases volatility or whether volatility invites more HFT – or both or neither” (ibid. p.65). The last remark summarises the state of the knowledge on HFT and highlights the need of further research. Moreover, even those effects over which some convergence among authors seems to exist, do not clearly show the same behaviour under both normal and market stress conditions. Since the focus of many studies on HFT is mainly on the impact it had on the Flash Crash, or that it may have in causing future crises, researching how HFT interacts with exchanges is of the utmost importance. Chapter 4 will focus on the difference between quiet and troubled market conditions.

2.2. LITERATURE REVIEW ON THE FLASH CRASH

2.2.1. INTRODUCTION

It is widely accepted in the literature, although with some exceptions, that the Flash Crash originated at the Chicago Mercantile Exchange (CME), namely in the E-mini S&P 500 futures contract market [CFTF-SEC (2010a), Kirilenko et al. (2011)]. This was one of the reasons for selecting the CME as the reference environment for the simulations (chapters 4 and 5) and for the data analysis (chapter 7). Yet, according to Menkveld and Yueshen (2016), the unusually large amount of futures contracts traded by Waddell & Reed (W&R) on the CME were not the cause, at least in a direct manner, of the Flash Crash. One of the sharpest criticisms to the CFTC-SEC reports comes from the market analysis firm Nanex, published in a series of articles on the internet. In the blog 'Zero Hedge', Durden (2012) goes as far as to rhetorically asking whether "was the SEC 'explanation' of the Flash Crash maliciously fabricated or completely flawed out of plain incompetence?". Reporting findings from Nanex, the article points out as "it was precisely HFT quote churning that was the primary, if not sole, reason for the catastrophic chain of events". At this stage, it is necessary to warn that despite Nanex reports cannot be categorised as academic research in the strict sense of the word, they are nonetheless a very authoritative source of information and dismissing them on that basis would not deliver a good service to academic research. Based on evidence, Nanex (2010b) "think[s] the delay in NYSE quotes was at the root of this detection", contrary to CFTC's findings. Nanex (2010a) arguments the above by showing evidence of the Consolidated Quotation System (CQS) not operating normally and within capacity by plotting CQS traffic at 2ms intervals versus the same plotted at 1 second intervals. The former message rate trespasses several times the 250,000 messages/second whereas the latter graphs misses to show that fallacy. A coarser granularity of the analysis would thus completely fail to notice the CQS delay, therefore deflecting the focus of the investigation toward other causes. On a more formal ground the

Foresight (2012), founding its arguments on a set of nearly 60 authoritative reviews, impact assessments, and papers, reaches the conclusion that "there is as yet insufficient evidence to what role HFT played either in the Flash Crash or other mini crashes that have occurred since HFT became established" (ibid. p.141). This uncertainty is shared by Kirilenko et al. (2011), who recommend further data analysis by making use of data from all venues, products and traders on May 6, 2010, in order to carry out an examination of Flash Crash hypotheses. They conclude the study with the rather intuitive statement that "irrespective of technology, markets can become fragile when imbalances arise as a result of large traders seeking to buy or sell quantities larger than intermediaries are willing to temporarily hold" (ibid. p.38). However, their neutral approach about technology sounds like an attempt to deflect Flash Crash responsibilities away from HFT, somehow indirectly confirming the findings of CFTC-SEC (2010a, 2010b) so harshly criticized by Nanex. By considering market fragility as not directly caused by HFT on the Flash Crash day, Kirilenko et al. (2011) implicitly suggest the former not being directly related to the latter. It could be instead a consequence of systemic risk, but according to Danielsson and Zer (2012), there is no clear consensus on what constitutes systemic risk. Indeed, lack of consensus on many important financial issues and even on some basic definitions is a recurring theme imperiously calling for settlement, if solutions to the problems financial markets are facing (the Flash Crash being only one of them) are to be found. However systemic risk is defined, Cliff (2011a) foresees it likely to grow rather than diminish in the next future, unless appropriate actions are taken. And although the probability of such events taking place is small, their potential consequences are so serious and so far-reaching in space and time, that appropriate actions are seen as urgently needed. In her testimony rendered to the US Congress on the Flash Crash, the SEC Chairwoman admitted that "the technologies used for market oversight and surveillance have not kept pace with the technology and trading patterns of the rapidly evolving and expanding securities markets" [Schapiro (2010), p.17]. To cure this weakness many authors have advanced sensible proposals. Among them, Madhavan (2012) regards the Large Trader Reporting rule, approved by the SEC in July 2011, as "an important step towards a consolidated audit trail that will give the regulators the tools necessary to monitor trading patterns across exchanges and improve enforcement" (ibid. p.23). Bullock (2011) highlights the emerging need for financial system simulation and Sornette and von der Becke (2011) go in deeper details by stating that "we need to build policy making devices (a 'policy wind tunnel' as Nigel Gilbert would call it, or an 'economic flight simulator')" (ibid. p.15). This research intends to bring a contribution in that direction. In the following sections several aspects of the Flash Crash will be reviewed. Some of them, like volatility and liquidity, which were already discussed in general in the HFT review, will now be analysed with respect to their relationship with the crash.

2.2.2. ANATOMY OF THE FLASH CRASH

May 6, 2010 shall be remembered by trading practitioners as the day of the Flash Crash. The market opened in a nervous mood because of worries of British Parliament being in a stalemate situation and of riots in Athens due to the austerity programme launched by the Greek government coupled with a BCE press conference that did not mentioned any purchasing of Greeks bonds. By 11am (time in New York) the Dow Jones was already down 60 points. At 1:30pm the VIX (Volatility IndeX) started to rise and the US Treasury yields to drop following a flight to quality by investors looking for a safe harbour. At 2pm the VIX reached 28.6 whereas the Dow Jones Industrial Average (DJIA) was down 161 points (-1.5%) and the S&P 500 -2.9%. At 2.23pm the NASDAQ started to issue alerts of abnormal behaviour by many securities and one minute later the first stock was traded against a stub quote, more than 80% below the previous day closing. The VIX trespassed 31.71% and did not show any intention to stop or slow down its climb. The financial firm Waddell & Reed started its heavy sell programme of E-mini S&P 500 June 2010 futures contracts. At 2:37pm NASDAQ declared self-help against NYSE Arca and at 2:45:28pm the CME, which was down more than 9.4%, stopped trading operations for five seconds. In the meantime the number of Liquidity Replenishment Points (LRP) hit the psychological threshold of 1,000 and over 200 securities had dropped 50% or more from their value just three quarters of an hour earlier. Despite the march to normalisation led by the CME, other markets kept behaving abnormally. The Dow hit its daily low and so did the NASDAQ. Some large capitalisation stocks traded at ridiculously low prices. At 2:49pm the BATS also declared self-help against NYSE Arca, which lasted five minutes, and one minute later 3M was put in LRP [source: CFTC-SEC (2010a)]. Further dramatic details are provided by Kirilenko et al. (2011). In the two minutes between 2.45pm and 2:47pm the Dow Jones Industrial Average (DJIA), the S&P 500 and the NASDAQ 100 reached their daily low; at 2:45:28 the E-mini S&P 500 futures for June 2010, already down to 1062 from 1165.75 at the beginning of the day, dropped 6 points in one second, down to 1056. The next transaction would entail a further 6.5 point decline and therefore, according to the CME Globex rules, it triggered the 5-second Stop Logic. After expiration of the halt period, the price fluctuated shortly and then started to rise, although with occasional declines, reaching 1122.75 at the end of the trading day, the same level it had at 2:35pm, well before the crash.

2.2.3. WHAT HAPPENED ON MAY 6, 2010

2.2.3.1. The View of the Institutions

A rather abnormal factor during the period of time later called 'the Flash Crash' was the selling program launched by W&R, which entered the market at around 2:32 and finished around 2:51. Its short position represented on average 9% of the volume traded over that period. As reported by CFTC-SEC (2010b), W&R sold on the way down and continued to do so even as the price level rose. The selling algorithm was programmed "to feed orders into the June 2010 E-Mini market to target an execution rate set to 9% of the trading volume calculated over the previous minute, but without regard to price or time" (ibid. p.2). The Securities and Exchange Commission labelled the analysis as 'hard'. The main challenges identified were the size of the analysis, with 17 million trades between 2pm and 3pm; and the complexity of the financial system, because of the interconnectedness of the exchanges, "diversity and opacity" (ibid. p.72) of the trading strategies, the linkages among different instruments and "the effects of computerized trading" (ibid. p.72). The possible responsibility of HFT on the Flash Crash has been raised since the day after the crash, mostly by practitioners and the general press. This was an expected reaction in case of inexplicable events. Yet, it makes sense to believe that those who, thanks to their speed, could afford it, tried to exploit the situation or at least to minimise the damage. As noticed by Kirilenko et al. (2011), at the beginning of the initial price decline the HF traders accumulated long net positions but they could quickly offset their investing before the price decline accelerated. Indeed, in the fifteen minutes before 2:45pm HFT activity became more and more aggressive, from being mainly neutral before the beginning of the crisis. Actually, HFT being harmful to markets, especially under stress, is a concept shared by many practitioners: this was the first hypothesis made in the aftermath of the crisis although it is intuitive to believe it was caused by a confluence of events, as suggested by Berman (2010). A communiqué from the Chicago Mercantile Exchange [CME (2010)] claimed no evidence that Algorithmic Trading (AT) caused the crash, nor any "undue concentration of activity" (ibid. p.3). Although the CME was obviously an interested party, their opinion is worth examining together with the others. The communiqué admitted inconclusiveness because of the quick analysis but boldly stated 'generally' balanced trading. Unfortunately the real issue at stake was not general balancedness but rather punctual unbalancedness, which was not denied, nor even merely mentioned by the CME. Their suggestion was therefore to look at "divergent trade practices and price protection mechanisms amongst the various stock trading venues" (ibid. p.4) for responsibilities. In her testimony before the US House of Representatives [Shapiro (2010)], the SEC Chairwoman also evenly distributed responsibilities to "a confluence of events" (ibid. p.5), substantially deflecting them away from HFT and also restated the (rather obvious) fact that "precipitous price declines are not exclusively associated with automated trading", implicitly discharging HFT from being the main culprit.

2.2.3.2. The View of Practitioners

A different view is taken by Durden (2010) who instead finds both unbalanced trading and concentration of activities in observing that in hundreds of cases stocks were quoted more than 1,000 times in a single second, with no economic underlying justification, and frequently at a price well outside the NBBO. It is a matter of fact that, despite all the public debates viewing HFT as the main responsible of the crash, no empirical evidence has written it on the marble. The early theories attempting to explain the Flash Crash included, together with HFT and the large sale by W&R, fat fingers (human errors), changes in market microstructure (market fragmentation and tick decimalisation), and technical issues (network delays, which were actually experienced). There is no doubt that at least some of these causes had a role in the event and many other have been proposed as, perhaps partial, contributors to the crash. It is true, though, that financial markets have suffered bubbles ever since; it is difficult to identify the first crash in market history, and it is also unclear whether high speed trading makes things any worse. In 1987 there were no HF traders and computers on the trading floor were rare - yet the stock market crashed. It was an unexpected event, a crash of that size was calculated to statistically occur once every several million years, instead it was there - and real people had witnessed it, and suffered from. The occurrence of financial bubbles has often been assimilated to the Black Swan [Taleb (2007)], a symbolic representation of highly improbable events that happen much more often than normally expected. The criticisms to the applicability of the Gaussian curve to real life events, and the financial ones in particular, can be grouped in the much mentioned 'fat tail' problem. Such a problem manifests itself when highly improbable events that should be confined in the remote ends of the bell curve are more frequent than foreseen by the Gauss' theory, resulting in not so thin tails of the curve. Example of black swans are discovered by Johnson and Zhao (2012) who find a strong correlation between sub second trading and the occurrence of (theoretically unlikely) financial crises. Whether or not a causal relationship of the former to the latter exists, in the authors' view it was a fact nevertheless. However, some even believe not just that the correlation does not exist, but it is in the opposite direction. By criticising market data-based systemic risk measures as unreliable, Danielsson and Zer (2012) view trading speed as a factor of market stabilisation, able to quickly cure temporary anomalies that, if left at the mercy of slow market waves, could degenerate into extreme outcomes. May 6, 2010 could be a good example of an extreme event that was quickly cured and the speed of recovery could mean that the fast participants had a role. What could happen if the market is too slow to recover after a proper, or mini, crash and, as Cliff (2011a) wonders, closes before the heal has a chance to be cured? The panic could propagate to the Asia-Pacific markets and then to Europe before Wall Street wakes up again to continue a recovery that, given the now dramatically changed market situation around the world, might never happen. As is common when opinions differ bitterly, a middle-of-the-road position also exists and the paper reports of "unforeseen dynamics of interactions between honestly-intended CBT trading systems" (ibid. p.5). The 'fallacy of composition' is also mentioned by Zigrand, Cliff and Hendershott (2012), who notice how the global outcome could be unstable even if all the system's components are stable on their own. This is a well-known result from the theory of complexity, an engineering discipline. If confusion was not enough, one has to bear in mind that analysis is also hardened by the culture of secrecy that permeates commercial tools supposed to be money machines. Another paper from the same author, Cliff (2011b), argues that regulators may lack the resources, or even the skills, to carry out appropriate analyses of disruptive events. Whatever the reason, these two factors make understanding just harder. Brogaard, Moyaert and Riordan (2014) agree with this view: the mechanisms which make prices to jump are not well understood and how HF traders behave during periods of price instability is little explored. It is not even completely clear if price jumps are liquidity-related or are a consequence of macroeconomic news. And although the halt mechanism seemed to work on May 6, Sornette (2003) suggests a more cautious approach: in October 1987 the countries enjoying the most stringent circuit breaker requirements also suffered the heaviest losses. Many other reasons explaining, at least partially the crash have been proposed: Easley, de Prado, O'Hara (2011) summarise a few, ranging from the unmissable fat finger to technical difficulties at NYSE and ARCA, from USD/Yen exchange rate movements to put option purchases, from the evergreen E-mini sale by Waddell & Reed to quote stuffing. By now one point should be clear: the causes of the Flash Crash are still unclear.

2.2.4. STUB QUOTES

The large majority of securities were affected by the crash. Yet, about 86% of securities reached lows for the day that were less than 10% from the 2:40 price. The trades executed at such discount were nearly 5 million. Many of the remaining securities experienced greater declines, with some trading at as low as one penny. Trading at such ridiculously low prices happened as 'stub quotes' were in place. Stub quotes are designed to formally meet the market makers' obligation of providing continuous two-sided quotation, but usually they do not dictate the price. So it happens that under extremely troubled conditions orders are posted at such low (or high) price they are not actually meant to be executed. Instead, some of them were and, as noticed by CFTC-SEC (2010b), their number represented a significant percentage of broken trades on that day. As soon as a market order, also known as 'at best' order, is submitted, it is immediately matched to the best available price on the other side of the book, even if the prevailing price is a stub quote. This is what happened in many cases on May 6. However unusual stub quote execution could be, "even more extraordinary was the fact that over 20,000 trades representing 5.5 million shares were executed at prices more than 60% away from their 2:40 p.m. value" (ibid. p.65).

2.2.5. WAS VOLATILITY CAUSED BY HFT?

As seen in the previous chapters, one of the most deeply studied issues about HFT is its effect on volatility. This aspect took even greater importance in the aftermath of the Flash Crash as not only many practitioners reported anecdotal evidence of the responsibility carried by HFT in the Crash but this (so far not demonstrated) hypothesis also matches common sense that only ultra-fast machines could cause such a swift price decline and rapid recovery. Vuorenmaa and Wang (2014) develop an agent-based simulation of the Flash Crash by having market making HF traders "collectively create a feedback loop system triggered by a large institutional sell, consistent with the widely cited 'hot-potato effect'" (ibid. p.1). This working paper also notices that tick size has an impact on volatility. The paper, referring to prior studies, argues that a larger tick size, decreasing the number of available order-book levels close to the mid-quote, increases the depth at each level, reducing volatility. On the other side, a larger tick size also increases bid-ask spread and a larger spread increases noise trading which is readily reflected in higher volatility. The critical question is deciding which of the two contrasting factors rules. The conclusion according to Vuorenmaa and Wang (2014), is that in 'normal' times volatility decreases with tick size but under stress conditions the probability of a crash is an increasing function of the number of HF traders and of the reduction of tick size. With 100 HF traders, the probability of a crash rises from about 5% with 0.1 tick size, to about 8% when tick size is 0.01 and about 13% if the tick size is 0.001. Whereas it sounds common sense that different market conditions lead to different behaviours, the inverse relationship between tick size and probability of a crash under stress is definitely a very interesting result. Among other considerations, it introduces the concept of 'phase transition' also discussed by other authors. By referencing papers that he wrote or co-wrote, in the Foresight DR7 [Sornette and von der Becke (2011)], Sornette suggests that in a world dominated by HFT, financial instabilities are expected to rise. His optimistic view, according to which "(i) the presence of bubbles can be diagnosed quantitatively before its demise, and (ii) the end of the bubble has a degree of predictability" (ibid. p.17) is balanced by the more neutral recognition of inability “to distinguish bubbles from time-varying or regime-switching fundamentals" (ibid. p.17). Even more pessimistically, Sornette and von der Becke (2011) refer to another couple of papers, one of which Sornette also co-authored, supporting "the evidence for the presence of bubble-like behaviours at arbitrary short time scales, that are often followed by strong corrections and swings" (ibid. p.18). Developing further the work on financial black swans in Johnson and Zhao (2012), Johnson et al. (2013) investigate the extreme events in financial markets with variable duration grouped by hundredths of milliseconds, e.g. 0 to 100ms, 100 to 200ms, and so on. They reach the enlightening conclusion that on the main US exchanges between 2006 and 2011 the number of black swans can be described by a near-exponentially decreasing function of the duration window. Also astounding is the result that spikes and crashes practically lie on the same curve. This result seems to suggest a close relationship between event duration and their frequency, once again fingerpointing HFT as the main culprit of extreme events. Colliard (2016) discriminates between traditional value-informed traders and a new category, called supply-informed traders, that possess no fundamental information but which know about the composition of orders, where the supply-informed traders are typically HF traders. One of the findings of the paper is that "supply-informed traders correct short-term mispricings: they behave as contrarian investors when the market overreacts, and as positive feedback traders when the market under-reacts to observed trades" (ibid. p.1). Thus, in this view, HF traders would mitigate volatility by trading in the direction of true (short-term) price.

2.2.6. WHY LIQUIDITY DISAPPEARED?

As noticed by Menkveld and Yueshen (2016) and Foresight (2012), the Flash Crash originated in E-mini contract market, one of the most liquid futures contract markets in the world. This is even more worrying, as it makes evident the potential fragility of the entire financial market structure. Asset fragmentation is investigated by Madhavan (2012), who finds that on May 6, 2010, it was significantly higher than in the previous 20 days. The study also finds "strong evidence that securities that experienced greater prior fragmentation were disproportionately affected" (ibid. p.3). Foresight (2012) summarises a few more illiquidity events in markets since the Flash Crash, like natural gas futures losing, and then bouncing back, 8.1% on June 8, 2011, or the five stocks listed on the LSE which experienced very rapid price changes on August 24, 2010. The Foresight Report concludes "evidence suggests that HFT and AT may be contributing to periodic illiquidity in current markets. They may not be the direct cause of these market crises, but their trading methods mean that their actions or inactions may temporarily amplify some liquidity issues" (ibid. p.57).

This conclusion is consistent with the main thesis stated by Kirilienko et al. (2011), according to which HFT may not have triggered the Flash Crash but it is likely to exacerbate the events in a period of market stress.

2.2.7. EVIDENCE

Research is not made only of econometric models but practitioners' opinions are also worth gathering. According to a survey conducted by Market Strategies International, whose results are reported by Kirilenko et al. (2011), more than 80% of US retail advisors believe HFT were the primary contributors to the volatility experienced during the Flash Crash and that a similar event could easily happen again. Yet, it is to be restated that, as seen in the previous section, this worry is shared more between practitioners and the general public than by academics. It looks like that, according to many rigorous studies carried out, evidence does not support worries about harmful impact of HFT on market stability but such studies are by no means conclusive. Kirilenko et al. (2011) interpret the large coefficient for aggressive mean reversion parameters as evidence that HF traders, which more often than not are posting passive orders, on May 6 quickly moved to the other side of the trades by submitting market orders, or marketable limit orders. Yet, an explanation could be that HF traders did not purposely change their behaviour during the Flash Crash but simply followed inventory-reduction strategies they tend to apply in every high-volatility situations for risk-reducing purposes. Indeed, the figures Kirilenko et al. (2011) find for HF traders about %-age volume, %-age of trades, average trade size, and average order size look similar to those of the three previous days but it must be highlighted that such figures are averaged over the entire trading day or days. Punctual behaviour might be, and during the Flash Crash actually was, rather different. Another strange behaviour, an event within the event, was the 'hot-potato' effect. During the fifteen seconds leading to the halt, HF traders bought and sold more than 27,000 E-mini S&P 500 futures contracts to each other. Kirilenko et al. (2011) attempt an explanation of this unusual effect by noticing that many non-HF traders had temporarily withdrawn, as they were either unable or unwilling to submit orders, which left no other participants except HF traders on the market. The net result was a marked increase of trading between HF traders. Yet, although this explains the rise in percentage of HFT-to-HFT trading, no reason for the increased frequency of this activity in absolute numbers is supplied. An explanation of the causes of the 'hot-potato' effect is attempted by Vuorenmaa and Wang (2014) who simulate the Flash Crash with an Agent-Based Model. The paper notices as tight inventory control is a key characteristic of HFT market making and concludes the reasoning by stating that "with an abnormally large selling pressure from an ALGO agent, HFT inventories are filled fast, making the hot-potato scenario more probable" (ibid. p.20). According to the Central Limit Theorem, tossing of N coins would approach a normal, or Gaussian, distribution as N tends to infinity, with 99.73% of outcomes lying within three standard deviations from the mean. However, as pointed out by Johnson and Zhao (2012), market crashes larger than 3 standard deviations occur much more often, concluding that “[t]his relative abundance of extreme behavior in the real world (e.g. stock crashes) as opposed to a coin-toss world, suggests that real-world systems represent the effective opposite of a collection of independent stochastic processes” (ibid. p.10). This assumption may lead to a profound re-consideration of the market crash investigation methods. Thus, extreme behaviour or, in general, volatility in itself does not seem to provide the bulk of financial information as much as fat tails, that is, deviations from the normal distribution. Indeed, that financial returns at the daily frequency exhibit fat tails has been recognized since the work of Mandelbrot and Fama in the Sixties of the last century, as mentioned in Danielsson and Zer (2012). Markets seem to follow a logic different from the one described by econometric theories and distributions do not seem to follow Gaussian curves. Supporting evidence to non-standard behaviour of the markets is also provided by Johansen and Sornette (2010), who find that most crashes have endogenous origin that is, self-reinforcing speculative bubbles, rather than being triggered by exogenous events, like macroeconomic news. Overall, the academic community admits that its knowledge of the cause of market crashes is far from definitive.

2.2.8. WADDELL & REED

The CFTC-SEC (2010b) report ('the Report') is quite bold in stating the responsibility of a single trade, 75,000 E-mini S&P 500 futures contracts sold in a short period of time, as the root cause. Among the 'Lesson learned', the Report lists the following: "Under stressed market conditions, the automated execution of a large sell order can trigger extreme price movements, especially if the automated execution algorithm does not take price into account" (ibid. p.6). This interpretation of the Flash Crash events is, at some extent, shared by Menkveld and Yueshen (2016), who admit that W&R "did not cause the steepest price declines in a direct manner" (ibid. p.2), as it contributed for only 4% of the total net sells, yet its trades seem somehow related to the decline. Closely inspecting order books in the minute before the CME Stop Logic mechanism was triggered, the paper finds that a chain of aggressive sell orders posted by other traders followed an occasional market sell order by W&R after 300 milliseconds. The interpretation is that the large seller indirectly affected the aggressive sell chain. The pretty obvious objection that after the Stop Logic the sale flow did not have the same disproportionate effect as it had before, is anticipated by arguing that W&R "did not find outside customers before the halt" (ibid. p.3). In order to support their argument, the authors state, by making use of a calibration exercise, that the price paid by the large seller was excessive. This has been indirectly confirmed by W&R itself. Indeed, the firm stated that the selling algorithm was programmed to target execution rate but without regard to price or time [CFTC-SEC (2010b)]. Several academic articles [Brewer, Cvitanic and Plott (2013), MacKenzie (2015), and others] pointed to this direction, yet, often taking it for granted by the authoritative source (the CFTC-SEC Report), whereas those studies which investigated the matter independently use to take different stances. The same fundamental seller had implemented a similar strategy in the past with no adverse consequence, either for itself or the market at large. The algorithm was programmed to minimise the price impact. Indeed, the 9% volume limit with respect to the previous minute was empirically considered a safe measure to prevent price decline, let alone the dramatic nose-diving that actually took place. Moreover, before the Stop Logic W&R was only able to sell less than half of the total amount of contracts, with the rest being sold during the price recovery period. The message from the CFTC and SEC was reassuring: the combination of unlikely causes their Report identified has an infinitesimally low chance to occur again. As seen above, things are not so clear cut, though. Moreover, whereas in the past the selling process terminated in around five hours, on May 6 it only needed twenty minutes to accomplish its task. If something was different, intuitively that must have been outside the large selling algorithm. On the other side, the strategy implemented by the algorithm, targeting a fixed execution rate without regard to price (or time), sounds weird and it seems likely that, once learnt this software design feature, regulators felt happy to have found the cause of the Crash. Actually, regulators were not the only ones: a study on the securities class actions by Levens (2015) highlights how "[i]ssues of reliance will likely prove insurmountable for HFT plaintiffs due to the traders' seemingly absent concern for the market price of the securities in which they trade" (ibid. p.1557). (According to the same study "[t]he current state of litigation involving HFT reflects the prevailing view that HFT are villains"). One of the main criticisms to the CFTC-SEC (2010b) report comes from Tyler Durden's blog (based on Nanex data and analysis). In fact Durden (2012) reports of an email exchange with one of the co-authors of Kirilenko et al. (2011), where he points out how the W&R algorithm only quoted limit orders, never crossing the bid-ask spread. The logic conclusion from Durden (2013) is that the algorithm "was entirely passive, and did not move the order price lower once entered. This type of passive algorithm cannot cause a market crash on its own". Similarly harsh is Cliff (2011b): "in the immediate aftermath of the report's publication it became clear that the CFTC-SEC version of events was at best contested and at worst directly contradicted by the 'tape' data" (ibid. p.6). The CME press release issued immediately after the publication of the Report, argued against its conclusions. A couple of weeks after the publication of the Report the SEC itself [Berman (2010)] made it clear that "it would have been premature to conclude that this [W&R's] trade played a leading role in the main event" (ibid. p.2). Pushing the observation further, McInish, Upson and Wood (2012) notice how not even the rapid increase of volume was due to W&R's strategy as it started well before 2:32pm, when the trade driven by the algorithm occurred, as identified by CFTC-SEC (2010b) and Kirilenko et al. (2011). In summary, it seems reasonable to say that, whereas the large sale of more than $4bn of E-mini S&P futures did certainly not contributed to brighten the gloomy feeling of the day, the investor community should be rather cautious before assuming that just avoiding such large sales would prevent another flash crash to occur altogether.

2.2.9. STOP LOGIC AND OTHER HALT MECHANISMS

CFTC-SEC (2010a) highlights another problem occurring in volatile markets: when an incoming order would result in trading at the level for which a LRP is set, the NYSE would stop accepting automated orders for that security, just to restart automated quotations as soon as possible. "In many cases, this occurs in a fraction of a second but when the market is particularly volatile, it can take a minute or more" (ibid. Appendix A, p.14). Most of the price decline occurred in the minute before the CME Globex triggered the Stop Logic [Menkveld and Yeushen (2016)] and after just twenty minutes there was no longer any visible trace of the most dramatic event of recent financial history. Amazingly for a short period during which any known rule broke down with dramatic effects, the 5-second Stop Logic triggered by CME Globex system seems to have worked as expected by giving the much-needed respite to the stressed markets, injecting the sought liquidity and restoring investors’ confidence. In those five seconds, the *κόσμος* reaffirmed its dominance over the *χάος*. Volumes traded ought not to conceal the fact that May 6 was an extremely illiquid day, as pointed out by Easley, Lopez de Prado and O'Hara (2011). In fact, the number of LRPs triggered on the NYSE substantially increased above average. The NYSE called LRPs on one thousand stocks, where a typical daily number is around 50. The abnormally high number of LRPs highlights the severity of market conditions and this evidence played a role in the decisions to reduce liquidity, pause trading or withdraw, taken by many investors.

2.2.10. FORCED SALES

Many authors point out that, by and large, HF traders mostly act as market makers, although not subject to market making obligations, by quoting limit orders. However, lack of obligations may allow HF traders to enter behaviours prohibited to traditional market makers. As noticed by Danielsson and Zer (2012), HF traders can quickly reverse their standing, especially during turmoil, and aggressively closing their positions. Therefore rapid liquidation, due to loss reduction strategies, may exacerbate a pre-existing downward pressure. Downward pressure cannot exist without sale. From this undeniable fact it may be derived the statement that if the pressure is large then sales must also have been large. This stringent logic is at the basis of the W&R hypothesis but it is by no means exempt from criticisms. Another hypothesis is that the large sale was distributed among several participants that, because of similar strategies or because of external factors, collectively created a cumulative large sale. One of the most important external factors has been identified by Leland (2011) in forced sales. Forced sales occur when an investor borrows in order to increase the funding available for investment. In such cases the exchange rules usually impose margin requirements: if the cash value of the assets held by an investor falls below a pre-determined fraction, the investor is demanded additional cash as a guarantee the debt will be honoured. If the cash requirement cannot be met immediately, assets must be sold to either satisfy the requirement or to reduce the cash margin due, or both. Under stressed conditions, the effect of forced sales can be self-reinforcing: "[f]orced margin sales affecting many investors with similar portfolios, can put intense downward pressure on market prices. The resulting price decline in turn reduces equity and further asset sales are required" (ibid. p.6). As the paper points out, the vicious circle generated this way may lead to crashes even without any new information. Drawing a comparison between the crises in 1929, 1987, the Long Term Capital Management (LTCM) crisis, the August 6, 2007, the 2007-2008 subprime crisis, and the Flash Crash, Leland (2011) finds "a common link between crises despite the observed idiosyncrasies: sizeable volumes of forced asset sales" (ibid. p.12). The leverage originating forced sales has its strongest impact when the whole market declines, as the margin call becomes a generalised problem of all the leveraged investors. Since forced sales are undistinguishable from fundamental sales, they contribute to the general feeling that the securities hit by heavy selling are so because of endogenous reasons, and not necessarily because of exogenous ones, like margin call. This makes the future appear gloomier than it should - a psychological effect is added to the purely arithmetical one. In fact, the theory states, and common sense suggests, that forced sales would impact less on price were they identified as such rather than, mistakenly, as informed sales. In all the crises considered in that paper, the speed at which investors forced to sell did so greatly surpassed the speed market makers could supply liquidity at, resulting in an imbalance that further reinforced the downward trend.

2.2.11. STOP LOSS ORDERS

If the role of forced sales was highlighted long afterwards by a third-party paper, the impact of Stop Loss orders was identified as a possible factor by the preliminary CFTC-SEC (2010a) report issued twelve days after May 6. Often without deep analysis, the impact role of Stop Loss orders in a declining market has been noticed by several authors. A stop loss order comes in association with a limit order, and it sets a price limit beyond which the position opened by the order would be closed with a still acceptable loss to avoid the worst. As soon as the price of the security trespasses the threshold indicated, the stop loss order turns into a market order and gets executed at the best possible price. When the price of the security moves fast, and in particular if it is moving during the time interval between the time the stop loss order is triggered and when it is actually traded, the stop loss order may execute at a price worse than anticipated. If at that particular price level there are several stop loss orders waiting to be triggered, their execution may unleash a series of consecutive selling (buying) that cause the market moving further down (up), potentially hitting worse and worse prices, which will be triggering other stop loss orders, and so on, in a fiendish, and possibly very long, chain reaction. The phenomenon just described can take even more dramatic dimensions if, as happened on May 6, liquidity is scarce. Stop loss triggering could result in "executions at aberrant prices" (ibid. Appendix A, p.A12). Its potentially dangerous effect is well known and it has also been recognised by the CFTC-SEC (2010a) report which also admits that "use of stop loss orders by other investors may have created additional sell pressure on ETF shares in a rapidly declining market" (ibid. p.50). In order to prevent the situation running out of control some exchanges set a limit to the delta between the order price and the corresponding Stop Loss price or prohibit market orders altogether, allowing only limit orders. However, use of Algorithmic Trading with little or no human control on individual operations to focus on speed of execution, risks to bypass the safety measure implemented by those exchanges. In these cases stop loss or market orders are converted into limit orders at prevailing price before being routed to the exchange. If the price moves adversely in the meantime, the limit order does not execute. At that point the computer recognises failure, cancels the limit order and immediately re-submits it, because the purpose of the algorithm is to minimise losses. This is what any human trader would probably also do. But in a fast moving market, this loop may occur several times, until the prevailing price reaches unreasonable levels. The CFTC-SEC (2010a) report acknowledges "an unusually large number of stop loss market orders" (ibid. p.30) during the Flash Crash but without providing any further evidence. Stop loss orders' contribution to troubles is not new. Cliff, Brown and Treleaven (2010) attribute 1987 Black Monday, at least partly, to "dropping prices hitting the trigger-points of these simple automated trading systems, and thereby causing them to sell. As they sold, their sales depressed prices further, thereby triggering yet other automated systems to sell, pushing the price even lower, triggering others to sell, and so prices spiralled rapidly downwards into freefall" (ibid. p.7). *Nihil sub sole novi*.

2.2.12. DELAYS

Ultra-fast computers producing data that are being sent on ultra-fast networks is given for granted since several years. But the receiving end of the communication channel must also be able to handle the data at the required speed. To cope with unexpected peaks of data traffic the queueing concept is widely used: the first to arrive will also be served first; the others will patiently wait. In technical jargon this approach is called FIFO, acronym of First In, First Out. But in a distributed world, where many computers carry out their own tasks independently and only interact with others at some specified time to exchange information, the ability to consume data in a timely fashion is paramount. This ability may be jeopardised by many kinds of malfunctions but even a perfectly working system may experience serious problems when the amount of incoming data is too large for being handled timely by the available resources. Indeed, heavy data traffic can at times exceed the capacity of market systems to handle it and result in accidental delays. This is also pointed out by Linton and O'Hara (2012). Outages were observed on May 6. Similarly, in early August 2011 high volatility took its toll from the trading servers at several large trading firms, including Goldman Sachs. Delays of any kind experienced by the exchange servers have been recognised as a problem by Haldane (2011): "Message traffic resulted in delays in disseminating quotes for over 1000 stocks" (ibid. p.13). Although CFTC-SEC (2010b) agreed that when the markets are nervous message traffic can worsen the situation because exchange servers may have difficulty to handle the data flows, it was unable to find any evidence supporting the hypothesis that delays in the CQS and CTS feeds triggered turmoil on May 6. The official conclusion of the Report is that such delays, although undeniable, were not the primary factor of the crash. Quite the contrary, Nanex (2010b) analysed the NYSE delays and concluded that bid prices disseminated by NYSE were lagging behind and therefore remained temporarily above the falling ask prices at other exchanges, although the timestamps of the orders were updated correctly. So, they looked perfectly valid orders whereas in reality the prices were obsolete by several seconds. In particular, in the two minutes between 2:44:45 and 2:46:29 NYSE suffered an average delay of over 10 seconds in disseminating quotes for as many as 1665 securities, with peaks of more than 20 seconds for 40 of them and an average of 5 seconds for the others. The delay experienced by NYSE caused several spread crossing, giving rise to arbitrage opportunities. Durden (2010) reports of up to 250 stocks being affected. In the US the information about the quotes is being transmitted to the CQS. When a trade takes place the information is transmitted to a different system, the CTS. Since the number of quotes is much larger than the number of trades, there is rarely a delay in the CTS; chances are much higher that if any delay occurs, the CQS will suffer from it. According to Durden (2010), many HF traders received an outdated, and relatively higher, bid price and a few milliseconds later a lower trade price, which they interpreted as a sudden drop. Thus, their immediate reaction was to sell stocks to exploit the downward momentum. This behaviour caused a short term feedback loop which depressed prices further. An official Nanex bulletin, Nanex (2011a), confirms this: "CQS was widely believed to have been operating normally and within capacity on May 6, 2010. We have found strong evidence this was not the case, and in fact, CQS was severely saturated and therefore delayed during the flash crash" (ibid. p.1). The bulletin adds an explanation for the wrong belief. Message traffic is measured at 5-second intervals, with the result that punctual delays are averaged out and therefore lost in the large multitude of non-delayed data. Nanex (2010b) finds the delay in NYSE quotes, together with lack of CQS capacity, the major cause of the crash. Golub, Keane and Poon (2012) highlight the issue of fleeting liquidity, which occurs when the best displayed quotations are cancelled before the CQS could disseminate the information. This way the liquidity observed by a local investor is different from the one visible to a remote one. Linked to liquidity providing is the consideration made by Nanex (2010b): when the number of quotes in a single stock exceeds 5,000 per second the whole concept of market orders is to be abandoned. At that rate chances are too high of executing at a price different from the one observed when the trading decision was taken. This leads to a degree of uncertainty difficult to combine with market stability. During her testimony, Shapiro (2010) made it clear that "[t]he exchanges and other trading venues have adopted highly automated trading systems that can offer extremely high-speed, or 'low-latency', order responses and executions. The average response times at some exchanges, for example, have been reduced to less than 1 millisecond" (ibid. p.10). Beunza, Millo and Pardo-Guerra (2012) notice how all models built to represent the functioning of the financial markets have validity only under well-defined conditions; strong volatility, lack of liquidity and data delays almost always fall outside that set of acceptable conditions. It is therefore unsurprising that facing abnormal conditions the markets responded in a way not predicted by any model. Their conclusion is that "if you have delays in market data, your models are not going to be trading on reliable information. You have to shut that [trading system] down" (ibid. p.14). This is exactly what did many market makers and other liquidity providers on May 6, 2010. The conclusion drawn by Easley, de Prado, O'Hara (2011) is crystal-clear: "This generalized severe mismatch in liquidity was exacerbated by the withdrawal of liquidity by some electronic market makers and by uncertainty about, or delays in, market data affecting the actions of market participants" (ibid. p.3).

2.2.13. INTERMARKET SWEEP ORDERS

An original view of the possible cause of the Flash Crash is described in McInish, Upson and Wood (2012), the impact of Intermarket Sweep Orders (ISOs). In the US, the Regulation National Market System (RegNMS) provides a set of rules for governing the functioning of the financial markets. In particular, the Order Protection Rule forces a venue receiving an order that can be executed at more favourable price by another venue, to pass it over to the latter. Yet, there is a special kind of order, the Intermarket Sweep Order, which overrules that obligation. ISOs do not require price check and re-routing for a trade through price for order execution, and therefore are considerably faster than non-ISOs. It is up to the trader to specify an order to be an ISO - and it takes on the full responsibility of its choice. McInish, Upson and Wood (2012) report an example provided by the SEC, which shows how ISOs can add to market instability by increasing volatility, widening spreads and reducing liquidity. ISOs occur rather frequently, as traders whose strategies are mainly based on speed of execution cannot afford the luxury of wasting precious milliseconds in checking the (fractional) best price through other venues. Golub et al. (2012) adding practical evidence to the theoretical example, analyse quotes disseminated by the CQS and trades disseminated by the CTS for securities listed at NYSE, NYSE Amex and NYSE Arca during four months between 2008 and 2010. They find that out of 5,140 either up- or down-crashes, 3,488 (well more than two-thirds) were ISO-initiated. As it may be the case for other market features seen above, everything seems to work fine under 'normal' condition, however defined. And as liquidity diminishes, the number of ISOs increases. This can be explained with the hurry to ensure own order to be serviced before someone else's is. It thus matches intuition that a feedback loop is likely to arise: reduction in liquidity calls for more ISOs; ISOs consume liquidity locally and the loops repeats. Not surprisingly, the number of ISOs increased on May 6. Similarly unsurprising is the fact that, according to McInish, Upson and Wood (2012), the use of ISOs contributed to destabilisation.

2.2.14. MARKET MICROSTRUCTURE

In order to try and understand the Flash Crash (as well as many other market occurrences), market microstructure is an invaluable tool. By 'market microstructure' it is meant a branch of finance which deals with the inner details of how financial markets behave as far as price formation, transaction costs, exchanged volumes, fee structures, liquidity supply, order types, information asymmetry, exchange rules and inventory models are concerned. Another common definition is "the study of trading mechanisms for financial securities" [Krishnamurti (2009), p.13] and according to the National Bureau of Economic Research (NBER) "[i]t includes the role of information in the price discovery process, the definition, measurement and control of liquidity, and transaction costs and their implication for efficiency, welfare, and regulation of alternate trading mechanisms and market structures" (ibid. p.14). The most fundamental assumption of the theory is that the price of securities does not necessarily reflect available information due to how real-world markets behave. If markets were perfectly efficient, the details of how they behave would be irrelevant, since prices would follow a random walk, as stated by Fama (1965). At the contrary, market microstructure recognises that trading behaviours affect prices and admits the possibility of short-term alpha and risk-free profit. Market making obligations affect how markets behave, and so do priority rules. As seen in previous sections, absence of continuous two-sided quote obligations for market making allows HF traders to benefit from quoting rebates without suffering in troubled periods. Conversely, size priority rules (larger orders being executed first) favours institutional traders, more likely to submit large orders, whereas price-time priority rules tend to favour faster traders. On the other side of the spectrum, pro-rata execution would make speed advantage much less relevant. Krishnamurti (2009) notices how liquidity can be interpreted according to more than one dimension, depending on the observer: it could stress the time taken to convert an asset into cash or the cost paid for cash conversion immediacy. Historically cost of immediacy has taken the greatest importance but in HFT times, speed of execution tends to carry more weight. Information asymmetry is also an important factor in market microstructure. The literature has categorised traders according to their informedness capability, where informed traders have an edge because of their fact-based trading, versus wishful-thinking crossing-fingers investors. Real Simple Syndication (RSS) are news designed to be read by computers rather than people. They have radically changed the scenario. News directly received, analysed, and acted upon by computers made the very concept of information-based trading a factor heavily affecting market microstructure.

2.2.14.1. Market Microstructure and High-Frequency Trading

The origin of microstructure theory can be traced back to the Seventies of the last century, when Garman (1976) modelled dealer and auction markets using a collection of market agents. Other studies followed occasionally once every few years, until the Nineties, when research on the topic started to flourish, until the theory was systematised by O'Hara (1995). The microstructure of the financial markets has gained further importance since the introduction of HFT. At very high speed even the minuscule details matter since any bit can make a large difference. This descends directly from the non-linear nature of HF markets. Market microstructure theory also claims to have an answer to the Flash Crash questions - and to the market turbulence issues in general. Since the beginning of the new millennium, HFT firms have steadily grown their activities on the markets, at least until a few years ago. CFCT-SEC (2010a) estimates HFT volumes to be often around half of the total whereas Cliff, Brown and Treleaven (2010), as well as Ahlstedt and Villyson (2012) and Foucault (2013), indicate more than 70% of the trading volume being executed by HF traders, although they are only representing a mere 2% of the 20,000 firms active on the financial markets. A more cautious number is provided by Friederich and Payne (2011), who only recognise 54% of the volume as attributable to HFT while the authoritative speech by Haldane (2011) acknowledged a consistent growth from less than 20% in 2005 to a variable fraction between 2/3 and 3/4 at that time. The rationale behind the growth unanimously noticed by the researchers is that HFT firms have the potential to earn very tiny margins on massive numbers of trades. This is true in particular when the order flows are essentially balanced. When things turn bad it is a whole different story. With unbalanced order flows market makers face strong chances of losses because of adverse selection. Such losses represent what Easley, Lopez de Prado and O'Hara (2011) call 'toxicity' and its estimation reflects itself in liquidity providers' participation. If the perceived level of toxicity becomes too high, market makers will liquidate their positions and likely withdraw: extreme toxicity may transform liquidity providers into liquidity consumers. In stressed markets, the absence of market making obligations enjoyed by HF liquidity providers, leads to sudden withdrawals, potentially resulting in heavy illiquidity conditions.

2.2.14.2. Market Toxicity

One of the sharp differences noticed by Easley, Lopez de Prado and O'Hara (2011) in a HF framework with respect to more standard microstructure models is the meaning carried by the concept of time. Since trades take place in a matter of milliseconds, trade time becomes the most relevant metric rather than clock time. Trade time can instead be measured in terms of increments of volume. The three scholars introduce a metric that represents the toxicity of the markets: the VPIN, the Volume-Synchronized Probability of Informed trading (VPIN is a trademark of Tudor Investment Corp.). Nobody better than them can explain it. "For any time period, the VPIN metric is the ratio of average unbalanced volume to total volume in that period. Heuristically, the VPIN metric measures the fraction of volume-weighted trade that arises from informed traders as the informed tend to trade on one-side of the market and so their activity leads to unbalanced volume (either more buy volume than sells volume or the reverse)" (ibid. p.5). It follows that when trading tends to be mostly driven by information, the VPIN metric will grow large. The authors compute the VPIN metric for the E-mini S&P 500 futures contract between 1/1/2008 and 30/10/2010 by calculating the VPIN for each period during which the traded volume equals 1/50 of the average volume for that day. Their finding is that the VPIN value for the securities under investigation rated abnormally above average for the whole week before the Flash Crash. On May 6th, by 11:55am the value of the VPIN metric was in the highest decile, shortly after 1pm it was in the 5% tails of the distribution to reach the highest value ever by 2:30pm, just two minutes before the crash started. The study shows a slow increment of the liquidity issue in the days preceding the collapse, indicating that the toxicity of order flow preceded, and somehow announced, the disaster, even though the W&R trade may have exacerbated a situation already on the verge of precipice. A factor that might have further contributed to the generalised madness of that day was the decision of some market makers to close their long positions by accepting the losses, to withdraw or to quote stub prices. But despite all the contributing factors, the main points raised by Easley, Lopez de Prado and O'Hara (2011) are still powerful: the VPIN metric is a useful predictive measure of absolute returns and the Flash Crash could have been foreseen by looking at the VPIN metric. More generally, the theory developed around the VPIN metric states that: "[a]t relatively normal levels, it is a measurement of flow toxicity [whereas a]t abnormally high levels, it can also be understood as indicating the likelihood that market makers turn into liquidity consumers" (ibid. p.12). The latter occurrence can also be interpreted as a warning measure for the risk of a liquidity crash. Order flow may become especially toxic when market makers are unaware they suffer losses while providing liquidity. In particular, market makers operating at high frequency usually do not make directional bets; they content themselves of extracting minuscule profits from a very large number of passive trades. The success of this HF strategy greatly depends on their ability to avoid being adversely selected; this means not being less informed than their market taking counterparts. Therefore the VPIN metric is a predictor of short-term toxicity-induced volatility. The authors conclude with two rather intuitive statements: [d]uring periods of unusually high VPIN metric values, we would expect some increase in stocks' volatility as a result of liquidity providers withdrawing from the marketplace" (ibid. p.12), and "[t]he 'flash crash' might have been avoided, or at least tempered, had liquidity providers remained in the marketplace. Not only did some withdraw, but arguably they became liquidity consumers by dumping their inventories, thus exacerbating the crash" (ibid. p.13). All this may sound common sense but in the paper it was supported by the in-depth and rigorous use of the VPIN metric, so making it scientifically strong and, in the authors' opinion "a risk management tool for the new world of high-frequency trading" [Easley, Lopez de Prado and O'Hara (2012), (ibid. p.1490)]. Needless to say, not all academics agree with the ground-breaking appreciation of the VPIN metric. Indeed, a fierce debate has developed between the authors and their critics.

2.2.14.3. An Academic Debate

A strong challenge on the value of the VPIN metric as a crisis predictor has been put forward by Andersen and Bondarenko (2014a). They carry out an empirical investigation on the same January 2008 through July 2010 E-mini S&P 500 futures contracts, aggregating transaction series into one-minute observations featuring the last recorded price and the total trading volume over that period. The result was that VPIN "is a poor predictor of short run volatility" (ibid. p.1) that was not displaying a peak before the Flash Crash, but only after it. Moreover, its predictive content is regarded as somehow implicitly included in the VPIN algorithm, being it mainly caused by the 'mechanical relation' with the trading frequency. They recognise that its value before the crash is high but by no means reaching a historical peak as it does afterwards. Other, more traditional, predictor variables seem more robustly correlated to short-term realised volatility than VPIN. At the contrary, VPIN is correlated with trading volume (whence its high value around the Flash Crash) and, "if anything, negatively related to future return volatility" (ibid. p.23). Obviously this sharp criticism would not go unnoticed by the original authors, who responded defending passionately the academic value of their findings, conceding that not all volatility is due to order flow toxicity (something they never affirmed) but stating that extreme toxicity can lead to volatility (by reducing liquidity) and finding a relationship between high VPIN and subsequent volatility. Moreover, Easley, Lopez de Prado and O'Hara (2014) remind how microstructure literature recognises trading volume as a critical variable to understand price movements, highlighting that this is because of its "linkage to underlying information dynamics" (ibid. p.48), essentially reaffirming the role of order flow toxicity and therefore of the VPIN metric. This would be particularly true in HF markets because market makers (mostly ultra-fast silicon traders) operate on a volume-clock rather than on a time-clock (as explained above). In fact, in order to take this into account, their model updates the VPIN metric every fixed amount of securities traded rather than every fixed amount of time. Andersen and Bondarenko (2014b) counter-responded to the above arguments, stating that VPIN is driven and distorted by traded volume and that since volatility displays strong time series persistence, it comes as no surprise that volatility possesses self-predicting capability. The debate does not seem closed, as yet. Even in this case, academics' opinions seem to diverge markedly, as seen above in many other areas of HF-related matters. Other HF features that affect market microstructure being investigated in a more recent article by Easley, Lopez de Prado and O'Hara (2016) are the practices of allowing (for a fee) HF traders to co-locate their computing resources to the exchange premises or to allow them seeing markets more clearly and a few milliseconds before ordinary participants receive consolidated tape data. This, at microstructure level, corresponds to turning public information into private information, which may be viewed by its critics as a form of legalised insider trading.

2.2.15. FLASH CRASH' LESSER BROTHERS

The Flash Crash spawned great worry among practitioners and regulators alike, and attracted a lot of interest by the academic world. Some argue that, leaving aside worries for the abnormal market behaviour, May 6, 2010 was, at the end of the day, just another troubled day on the financial market as many other were observed in the aftermath of the 2008 subprime crisis. Other researchers think differently: that event was by no means 'business as usual'. Surprising as it may sound, on May 28, 1962, nearly forty years before HFT was even invented, the markets experienced an *ante litteram* flash crash. One may decide whether to feel relieved, as HFT has nothing to do with such extreme events, or to worry about the implicit capability of the markets to display extreme behaviours that today's trading speed can only make more dramatic and more frequent. Indeed, several sources talk about mini flash crashes, rapid events occurring at sub-second scale, affecting only one or very few securities, much less noticeable than the May 6 event. Still, in the view of Vuorenmaa (2014), they are relevant for market stability and investors' confidence. According to Sornette and van der Becke (2011), mini flash crashes seem to happen rather frequently and, like the events occurred on April 28th and September 27th, 2010, they seem to take place in correspondence with an increase in quoting frequency. The market data provider Nanex identified thousands of mini flash crashes over the last few years. If a mini flash crash affects only one security, as suggested by Golub, Keane and Poon (2013), the event is much harder to detect than one displaying a world-wide multi-market contagion. Exactly as the major Flash Crash, mini events can raise serious doubts about the stability of the financial markets. Recent reports highlighted as this phenomenon is in no way restricted to the US market but it seems to affect the whole world's financial community. Because of the extreme speed at which things happen, lacking a thorough understanding of the causes leading to minor or major flash crashes, real-time human supervision would not be able to help at all, leaving the financial markets at bay of events whose unpredictability only relates to the 'when' rather than the 'if'. Instead, a self-reassuring position is taken by Gomber et al. (2011), by fingerpointing the US market structure. The report commissioned by the Deutsche Börse, states that, since Europe enjoys a more flexible ‘best execution’ regime thanks to the Markets in Financial Instruments Directive (MiFID) and a circuit breaker regime based on individual securities, “no market quality problems related to HFT have been documented so far” (ibid. p.58). Another European report, the Foresight (2012) commissioned by the UK Government Office for Science, keeps a more balanced position. Maybe because it has been published about one year after Gomber et al. (2011) and more evidence was by then available, Foresight (2012) recognises that “[t]here has been a variety of other, smaller illiquidity events in the markets since the Flash Crash” (ibid. p.57). The report lists the 8.1% natural gas drop on June 8, 2011, which bounced back in a few seconds; the eightfold spike in volatility of oil futures on February 2, 2011; the 98% fall in Morningstar ETFs in March 2011; the very rapid changes in BT Group, Hays, Next, Northumbrian Water Group, and United Utilities Group, all listed at the London Stock Exchange. In all these cases, no significant news seem to have caused the swings, which in some cases affected European market as well. But the list does not terminate here. Zervoudakis et al. (2012) mentions the Dow Jones Industrial Average flash crash on September 29, 2008, the cocoa futures mini flash crash on March 1, 2011 and the dollar-yen sell-off fifteen days later. On May 2nd, 2011 it was the turn of gold to drop by $20 just to quickly recover more than $15; silver followed suit on the next day. In July the same year crude oil futures showed large swings [Cliff (2011a)]. Recent years have witnessed other similar events. Sornette and von der Becke (2011), referring to a few other studies, summarises that “[m]ini-flash-crashes in single stocks seem to happen rather frequently” (ibid. p.13) and although the definitive proof of HFT involvement is missing, some crashes seem to have been accompanied by increase in booking frequency. According to Nanex, the last few years experienced thousands of mini-crashes. Investigation of the number of both crash and spike black swans in 100ms windows in Johnson et al. (2012) shows that such number increases as time between ticks shortens, with near-perfect superposition of the upward and downward curves. In particular, within the 100-200ms window the number of events was about ten times greater than in the 900-1000ms window. The authors also find 18,520 black swan events (defined as ten down- or up-ticks in a row with more that 0.8% price change) that lasted less than 1.5 seconds on multiple exchanges between 2006 and 2011: that is nearly 10 per every working day! Foresight (2012) suggests that a high number of mini-flash crashes may be caused by the feedback loop generated solely by computer algorithms. A possible reason for such short events to be neglected is that they tend to cure themselves nearly as quickly as they arise, in a matter of milliseconds, and are often assimilated to market noise. As seen before, several studies find that HFT tend to improve liquidity at normal times but some recognise a different behaviour and a different effect under stress. This seems confirmed by Golub et al. (2012), who analyse mini-flash crashes in the US equity markets during the four most volatile months in the period 2006-2011. Their findings confirm the adverse impact of HFT on liquidity during the mini-flash crashes. Moreover, analysing quoted liquidity during the same events, they note a stronger reduction on the bid side by HF traders, resulting in sell-side pressure, something that matches both common intuition and practitioners' perception.

2.2.16. CONCLUSION

The main debate about the causes of the Flash Crash is whether W&R, which initiated an unusually large selling program a few minutes before the crash, carries any responsibility at all over the occurrence of that dramatic event. Similar to other issues on the HFT topic, many different views co-exist although in this case the evidence contrary to fingerpointing one trading firm brings strong arguments and seems to carry a lot of good sense. Once again, easy explanations seem to have no citizenship in the HFT debate. Other paths need to be investigated in order to properly understanding the causes of the sharp plunge in prices. In fact, this paragraph analysed several other aspects and possible causes of the Flash Crash, including market microstructure, exchange delays, forced sales, Stop Loss orders, and the use of ISOs.

2.3. LITERATURE REVIEW ON MARKET TIERING

2.3.1. INTRODUCTION

In the academic world, with a few remarkable exceptions, there are not many studies about markets splitting into tiers; when the issue arises, in most cases it is a by-product of other reasoning. Even the exceptions focus mainly on fast traders dealing with each other without investigating how slow traders behave in those situations. De Luca et al. (2011) list five existing studies experimenting human-robot interaction in a financial market replicated in a laboratory. The main purpose of all the papers was finding out whether the robot did outperform human traders (which they did) rather than investigating market tiering. Cartlidge and Cliff (2012) carried out an experiment involving both electronic and human traders with the purpose of investigating how humans and robots interact. They find a statistically lower than expected level of mixed interaction between human and robot traders, which suggests some form of market tiering. During their investigation on the impact of low-latency activity on market quality using order-level NASDAQ data, Hasbrouck and Saar (2013) argue that, when operations occur at very short timescales, in the order of milliseconds, there is simply not enough time to analyse financial data going beyond the very local environment. This may lead to surmise the existence of one global marketplace, where investors from everywhere meet and trade, and a number of local markets in which local investors only look at what happens in their courtyard, simply because they cannot afford the time to watch elsewhere - so missing opportunities. The two scholars recognise that algorithms which submit and cancel quotes in the order of magnitude of milliseconds or below, target their similar fellows, implicitly setting the framework of two separate driving lanes: some sort of motorway and a footpath.

2.3.2. IN SEARCH OF PHASE TRANSITION

A source of worry is whether human controllers can apply good judgement and act upon activities which take place at a pace well beyond their capability to even understand what is going on. In this sense the Foresight (2012) report states that "an important speed limit has been breached" (ibid. p.85) and a phase transition has occurred. The report stresses the point: today's and, even more, future's market will less and less look like the 'same old movie' played at fast-forward speed. They are different financial systems altogether with respect to sensitivity, information and risk, as a quantum leap has occurred. Something different happens, phase transition seems to occur, market information is virtually impossible to match between different venues and the occurrence of feedback loops increases dramatically. Feedback loops in particular are extremely worrying because of their self-reinforcing effects, leading to potentially disastrous consequences, and because of the little understanding we have gained about them so far. This is the view carried by Cartlidge and Cliff (2012), who state that "[a]t sub-second timescales, below the robot transition, the robot-only market exhibits 'fractures' - ultra-fast swings in price akin to mini flash crashes - that are undesirable, little understood, and intriguingly appear to be linked to longer-term instability of the market as a whole" (ibid. p.3).

2.3.3. SUB-SECOND TRADING

The expression 'sub-second timescale' is quantitatively evaluated by Budimir and Schweickert (2009) in comparison to what is arguably considered the quickest physical action humans are capable of: blinking. Events lasting less than this threshold are considered irrelevant to human judgement. Most other human actions require considerably longer. The transition between events lasting more and less than such a threshold carry interesting properties. Using Shanghai market data recorded at millisecond level by Nanex, a financial analysis firm, Johnson and Zhao (2012) notice how at such scale fundamentally different regime of financial market behaviour emerge. They argument their rejection of the view of the 'same old movie' just played faster [as mentioned in Foresight (2012)] based on the observation that the near-universally accepted postulate of scale invariance self-similarity of markets no longer holds at ultra-high speed. "To a reasonable approximation, the patterns observed over months are similar to those over weeks, which are similar to those over days etc. We find that this is not the case as one moves through the sub second time barrier beyond which only machines can operate" (ibid. pp.3-4). This happens partly because of the algorithms' simplification at that scale, and partly because of the fat tails due to the likely knock-on effect on price volatility described therein. The sub-second crowding effect also impacts on the overall HFT profitability. The more HF traders operate in one market and the less profit each of them will have to content itself with. Brogaard (2011) states the existence of "some upper bound on how much market activity can be [done] by HFTrs before they are simply trading amongst themselves" (ibid. p.5). This view is agreed upon by De Luca et al. (2011) who plot the relative frequency distribution of trade executions for a set of experiments with slow and fast players implementing different strategies. They find that fast agents are able to a large extent to trade with each other before humans can intervene with their own orders. In such low-latency environments, the old contest between market makers and market takers re-proposes itself. Menkveld and Zoican (2016) produce a mathematically-founded theory for a scenario with HF market makers trying to cancel their limit orders before they become stale and quoting them again at the updated price according to the very latest news. On the other side, the HF aggressive traders try to pick them off by improving their own speed of reading and analysing news. Again, this scenario posits HF traders playing mostly against other HF traders. The practice adopted by many markets to sell information via direct data feeds to investors, allowing them to act on data before ordinary investors have received it, besides the already mentioned criticisms it attracts, also likely contributes to create a two-tier market. The informed ones will act on the news, either updating limit orders or crossing the spread, and the rest of the world will have to rely on official ticker tapes, delaying profit-grasping to a later time, if still available. Once again, speed seems to lead to phase transition in the marketplace. This view is somehow confirmed by Foucault, Hombert and Roşu (2016), who compare optimal strategy adopted by an informed trader when it can operate ahead of incoming news versus when instead it cannot. When traders act fast, they grab a larger fraction of trading volume and their trading is more correlated with short-run price changes. The conclusion drawn by Hasbrouck and Saar (2013) leaves no doubt: "it is clear that an algorithm that repeatedly submits orders and cancels them within 10ms does not intend to interact with human traders" (ibid. p.15). Bhupathi (2010) also rises the issue of a two-tiered market on the basis that HFT technology is only available to firms able to invest considerable amount of money - and to stand the arms race of never-ending technological update that follows the initial investment. Another, even more fundamental indication of the existence of a two-tier market is provided by the failure of time-scale invariance at sub-second scale. The fundamental point that Johnson and Zhao (2012) make is that below the second time-scale human traders, no matter how attentive, can no longer intervene. The reaction to any event in this timeframe is in the realm of machines only. They study the occurrence of extreme events, the main result being that the number of extreme price movements versus the clock time between ticks drops by about one order of magnitude when the duration passes from the 100ms to the one-second timescale and by another order of magnitude above two seconds. The same principle is also shared by Haldane (2011), who notices how "[n]on-normal patterns in prices have begun to appear at much higher frequencies" (ibid. p.9). Proceeding further from previous studies, Johnson et al. (2013) use a dataset composed of a millisecond resolution stream of prices for multiple stocks across multiple markets in the 2006-2011 period. They find 18,520 events fulfilling the definition of 'extreme' one (more than ten per day). Phase transitions seem to happen quite frequently in sub-second finance. According to Cartlidge and Cliff (2012) the final effect seems to be "longer-term instability of the market as a whole" (ibid. p.3). Davis, Van Ness and Van Ness (2014) study price clustering over a database consisting of 25 HFT firms in 120 stocks with different capitalisation on the NYSE and NASDAQ. They find that price clustering (the tendency of prices to occur on round or half-price measures) is less frequent when HF traders sit on both sides of a transaction than when low-frequency traders are involved. Whether positive or negative, this phenomenon marks a difference in the behaviour of HF traders. Phase transition seems thus confirmed by several studies.

2.3.4. CONCLUSION

All the worrying facts about HFT discussed so far should not lead to hasty conclusions. Of course, not everyone agrees that the market splits into two layers and some believe, on the contrary, that HFT is just the last step of a continuous process of search for speed and proprietary positions. One of the advocates of this party is Markham (2015), who concludes by stating that "HFTs are simply a continuation of market advances" (ibid. p.2). The same competitive advantages have been used before and for the same purpose; only the type of technology differentiates them from earlier attempts in that direction. Therefore according to this paper, any attempt to slow HFT down would be misguided and harmful to the functioning of markets. The debate is still very open.

2.4. LITERATURE REVIEW ON ARBITRAGE

2.4.1. INTRODUCTION

As far as arbitrage is concerned, the main focus of interest from the investors' point of view is the classical statement of the Efficient Market Hypothesis, developed by Eugene Fama in two seminal papers, Fama (1965) and Fama (1970). There exist three forms of the EMH: the Strong form states that the price of securities includes all the past and present information, whether publicly available or otherwise; the Semi-Strong forms claims that prices only reflect all publicly available information, whereas the Weak form further restricts the range of information contributing to price formation to the past. There is a wide consensus among economists, Grossman and Stiglitz (1980) among others, according to whom the Strong form of the Hypothesis does not hold, not least because insider trading (that is, trading on private information) is an illegal practice in many countries with financially advanced institutions. The debate, instead, is rather fierce on whether the Semi-Strong and the Weak forms have validity at all. There exist several wording for the Hypothesis but it can be summarised as stating that, assuming participants' rational expectations, in an efficient market all the prices do adjust instantaneously to any new information [Aldridge (2010)], preventing any market operator to consistently make abnormal profits. And, as argued by Fama (1970), who makes use of some theoretical models ('Fair Game', sub martingale, and Random

Walk models) as well as of DJIA data for the 1957-62 period, should abnormal profits in practice occur, they would disappear in presence of even minimum trading costs. Fama (1970) also prevents another criticism to the EMH, stating that some price movements, like large daily price change following a previous day large change would, if at all, contradict the Random Walk model but not the Hypothesis, in as much allowance must be made for some price oscillation due to the time necessary for correctly interpreting the original price swing. A few points are to be stressed about the EMH definition above. First: timing is critical. Then, the type of information also has importance. Lastly, by definition, consistent abnormal profit clashes with the very same concept of market efficiency. In the following sections, after some considerations on timing and profits, various types of arbitraging will be illustrated. Transaction costs shall also be discussed because of their impact in making markets more efficient.

2.4.2. TIMING

In the context of market arbitraging, timing is critical: if prices do adjust instantaneously, then it seems impossible to make abnormal profits in a consistent way from past (weak form) or even from current (semi-strong form) publicly available information. It is interesting to notice that the delay prices need to adjust to the newly arrived information is a function of the environment and it comes as no surprise that earlier authors considered compliant with the Hypothesis' requirements delays ["prices of individual securities adjust very rapidly to new information", Van Horne (1995), p.52] that more recent studies no longer do. This is a sign of technological advance, the rationale simply being that something ‘very rapid’ in 1995 did not allowed the time to exploit arbitrage opportunities, whereas the same very short delay turned out to be no more short enough fifteen years later, as in Aldridge (2010). The very concept of rapidity has changed its scope a few orders of magnitude over the same period.

2.4.3 CONSISTENT AND ABNORMAL PROFITS

According to the EMH, it is certainly possible to make profits by trading on the market but it is impossible to make abnormal profits (i.e. profits above the market average, or average profit running lesser risk) consistently without falsifying the Hypothesis. As a matter of principle, no trading strategy should be able to gain more profits than a naïve buy-and-hold strategy, at least in the long term. Fama (1965) takes as an example the net returns of 39 US funds over the 1951-1960 decade and notices that “[t]he most impressive feature […] is the inconsistency in the rankings of year-by-year return for any given fund” (ibid. p.92). Therefore, he highlights that "consistently is the crucial word here" (ibid. p.40). He also goes as far as to state that should a superior return be made, it would not necessarily be an evidence of superior knowledge, since "it is only when a fund consistently does better than the market that there is any reason to feel that its higher than average returns may not be the work of lady luck" (ibid. p.92). Fama (1970) affirms that the EMH derived from "accumulation of evidence" (ibid. p.389) of a path well approximated by a random walk. Another dimension of the analysis involves trading strategies; those strategies that, according to the Hypothesis, should never be able to consistently make abnormal profits.

2.4.4. EVENT ARBITRAGE

As is well known, news influence the market courses. There are many kinds of news that do so. They range from companies' periodical financial reports to macroeconomic news, from interest rates updates to commercial agreements, and so on. Many of the news on the headlines have the potential to make an impact, large or small, on the securities market. Classic financial theory told us that in an efficient market, prices adjust to new information instantaneously following a news release. In practice, market participants form expectations about announcements well before the official figures are made public. Aldridge (2010) shows the number of persistent trading opportunities in USD/CAD currency conversion in the period January 2002 through to August 2008, following an announcement about inflation rates in the US. Estimating results at 5-minute frequency, the announcements look providing no trading opportunities, at 1-minute frequency only 1 opportunity is detected, 5 are identified at 30-second frequency and 11 at 15-second frequency. The research does not go any further but the trend seems clear. News interpretation is a rather uncertain activity and in order to reduce the risk intrinsic to event arbitrage, a thorough analysis of how prices reacted to similar news releases in the past is usually called for. Even in this case, speed can greatly assist. The more time spent on analysing data and the more reliable the reaction, but the faster the reaction and the more profitable the trading. A trade-off between thorough, lengthy analysis and fast reaction is looked for, the ideal world being made by ultra-high computing power applied to large amount of data coupled with ultra-fast networking. Lacking either will lead to sub-optimal profit reaping, reaffirming the need of increasing the number of trading if substantial profit is to be gained. High number of trading per time unit equals to high frequency of trading. Market participants interpret incoming news according to individually developed parameters, which are reflected in the price they quote. The booked quotes provide other participants with information about valuation criteria. Then these other participants develop their own valuation, based on both the news and the prices just quoted by the previous readers. This process converges to a range of acceptable prices; the equilibrium price band can then be considered achieved. However, in presence of HF and non-HF traders, chances are that the former ones are the first to quote after the news release, whereas the latter ones have to content themselves with taking or leaving the updated price, or being picked off by faster and more informed aggressive traders. In all the arbitrage trading occurrences, the speed of booking a price in response to an event is likely to determine the trade profit or loss. In conclusion, event arbitrage opportunities suit well for being exploited by HFT operators and are most profitably executed in fully automated trading environments [Aldridge (2010)]. Mitra et al. (2016) anticipate advances of technology in the next future as far as faster and faster news collection and automatic interpretation is concerned, as a path those participants that already enjoy a speed plus. News Analytics technology is likely to be further developed "to automate filtering, monitoring and aggregation of news" (ibid. p.15), reducing latency.

2.4.5. INTER-MARKETS ARBITRAGE

Arbitrage between markets is a strategy conceived well before HFT entered the scene but it does require fast networks to retrieve timely information about prices for the same security from different markets, and as such, it is particularly suitable to HFT speed. If and when securities price discrepancy on different markets is noticed, then the standard operations of selling the security in the highly priced market and buying the same security in the other market applies. If both legs of the trading are executed with sufficient low latency to beat on time other investors, then a risk-free profit will be made. Trading speed has a definitive value in interconnected markets: "Multiple venues executing trades also means that prices need not always be the same, opening the door for arbitrage across markets" [O'Hara (2015)]. Aldridge (2010) notices that such a low latency strategy has a value in its own right. It is therefore sensible to assume that such kind of arbitraging is heavily used by HF traders.

2.4.6. INSTRUMENT ARBITRAGE

Many securities have closely related siblings in the same or in other markets. Typical examples are stock and the derivatives for the same stock. Up to a certain extent, and taking into considerations all the differences between different instruments, such securities tend to be somehow correlated. If the spot price of a security drops, under certain conditions it is likely that the price of the related futures follows closely. If the reason for the fall is perceived to have a long-term effect, then the futures price must adapt swiftly. Otherwise, if the consensus is for a temporary effect, the operators being long on such security will stand still, simply waiting for the consequences of the effect to pass by. But at that point, the price will have no reason to drop in the first place, or an arbitrage opportunity would be created, this time on the stock rather than on the futures. Aldridge (2010) acknowledges the existence of a lead-lag effect between the price of an equity and the price of the corresponding futures. At the time of her writing, and with respect to a research performed twenty years earlier by Stoll and Whaley, she recognises a dramatic reduction in such a delay, with stock-derivatives realigning adjustment dropping from between 5 to 10 minutes down to 1 to 2 seconds.

Yet, her conclusion is that, in presence of HFT systems that enjoy low transaction costs, this otherwise dramatic delay reduction is still insufficient to reaffirm lack of arbitrage-led profit-taking opportunities, and thus market efficiency.

2.4.7. RELATED SECURITIES ARBITRAGE

Arbitrage exploiting the relationship between different securities is based upon observation and discovery of pairs of securities that historically show similar patterns. The rationale behind such similar behaviours is to be deeply investigated to ensure that a rationale does exist, rather than being pure fortuitous coincidence. But once it is proved as statistically significant, then any divergence of one security with respect to the other is to be acted upon swiftly. Going long in the under-priced security and short in the overpriced one, is a standard arbitrage technique already observed in all other types of arbitraging activities. And, assuming that the historical similarity tends to be complied upon again, the investors who have got there first will, as usual, reap all the benefits. However, it must be highlighted that this type of arbitrage is not totally risk-free, as there is no 100% guarantee that the relationship between the two securities will reaffirm itself in the future.

2.4.8. STATISTICAL ARBITRAGE

Statistical arbitrage is close to technical analysis. Its goal is digging through large volumes of past securities data to devise undiscovered relationships - or to exploit known ones before others do. Whenever such relationship is violated, the statistical arbitrageur will assume that the price for that security will revert to the value predicted by the relationship, and will place its trade accordingly. Typical examples of relationships could be the historical high and low price of a security or the spread between two or more different securities, an extension of what has been described above in the section on arbitrage between related securities. The 'January effect' [Van Horne (1995)] is a well-known example. Another statistical arbitrage technique exploits historical behaviour of securities that, in absence of dramatic events, tend to oscillate within a fixed number of standard deviations from their mean. In particular, the Bollinger Bands technique allows investors to monitor securities' price level and volatility over a specified period of time. They depict stripes around the price bars, on the y-axis, versus time, on the x-axis. The price moving average line is sometimes also depicted. The stripe marks the distance of two standard deviations from each price point, where standard deviation is a measure of the security volatility. Therefore, the line describes the price and the height of the band takes into account the market condition corresponding to the price at the time. Whenever the price bar crosses the band, meaning it has oscillated more than two standard deviations, a signal will be generated. However, the signal might be a trend continuation or a mean reversal being ahead. In order to discriminate between the two opposite signals a direction indicator is required. The Relative Strength Indicator and Moving Average Convergence/Divergence are two among the most widely used signals by practitioners. The proper adoption of the technique, although not overly complicate, still requires a large amount of data to be processed and, according to the number of securities being monitored, a suitable amount of computing power.

2.5. CONCLUSION

Opinions on the impact HFT has on the markets are wide apart. For nearly any of the aspects analysed there are rigorous academic voices recognising an improvement of such parameter and others, not less authoritative, endorsing the opposite view. It must be said that scholars recognising a positive impact of HFT on markets seem more numerous but some partisans of the former side, like Kirilenko et al. (2011), switch camp when studying impact in troubled times. Similarly split are the interpretation of the causes of the Flash Crash. Institutional reports [CFTC-SEC (2010a)] and academics [MacKenzie (2015)] are quite bold in identifying in one single, albeit large, operation the root of every evil happened on that day, whereas the opposite stance [Nanex research or Easley, de Prado, O'Hara (2011)], points to diverse directions, ranging from technical problems to quote churning, from market microstructure to failure of well-established financial theories. Inspection of real data and complex mathematical models have so far failed to say the final word. Some researchers, for example Aldridge (2010), also find arbitrage affected by sub-second trading but this view is not unanimous. This research aims to bring a contribution to a deeper understanding of the impact of HFT on financial stability by using simulations, theoretical models and audit trail data analysis.

3. THE RESEARCH

3.1 DESCRIPTION OF THE RESEARCH

3.1.1. INTRODUCTION

At this stage, it is worthwhile to briefly summarise the findings of academia and the opinions of practitioners in order to distinguish between what is known for certain, what is likely, and what is still more or less in the dark. As seen in the Introduction chapter, there is no unanimous agreement on either the definition of financial stability or High-Frequency Trading. Therefore a research focussing on those two aspects will have to live with a certain degree of uncertainty. Nevertheless, there are some guidelines for either concept. Those situations falling in the middle of the definition are widely accepted; the frontier cases are more debated. For example, most trading days are deemed rather stable whereas May 6, 2010 was certainly affected by high financial instability. Less extreme situations are more difficult to assess as belonging to either category. On the other side, many academic studies find a beneficial effect of HFT on the markets: according to them [Brogaard (2010), among others] it improves liquidity while keeping volatility under control and leads to a reduction in transaction costs and bid-ask spread. However, this view is by no means shared by all academic researchers: some, for example Zhang (2010), find opposite results as far as volatility is concerned, and argue that a raise in volatility is likely to bring, under certain conditions, a reduction of liquidity. As reported by Gurkaynak, cited by Sornette and van der Becke (2011), "for each paper that finds evidence of bubbles, there is another one that fits the data equally well without allowing for a bubble" (ibid. p.17). Moreover, some academics highlight how current technological innovation solely based on sub-second speed advantage hardly brings any benefit to the society at large. The opinions of practitioners are less diverse. The voices of those opposed to the HFT practice (the majority among trading professionals) tell stories of uncontrolled volatility, ghost liquidity and frequent mini flash crashes, leaving room to the worst (albeit largely still-to-be-proved) theses available on the specialised and general press about maliciously created crashes and bubbles. After one Flash Crash, several mini flash crashes and an impressive amount of research on the topic, we are still unable to properly understand whether HFT can cause crashes or bubbles, even after it was shown one has occurred. The rapid pace at which they currently take place and their increasing number are no help: quite the contrary, the large majority of investors feel more and more disoriented as new articles get published on the topic. The literature on HFT and the Flash Crash is extensive, indeed some topics are investigated in more depth than others.

1. In particular, the impact HFT has on volatility has been studied inside out [among others, by Brogaard (2010), Zigrand, Cliff and Hendershott (2012), Myers and Gerig (2014)], yet the findings are not all pointing to the same direction: such difference of opinions suggested to research this topic even further.
2. On the contrary, most results found on market tiering [De Luca et al. (2011), Brogaard (2011) Cartlidge and Cliff (2012)] seem by-products of research investigating other topics and therefore it can be said that the matter requires deeper and more focussed investigation.
3. As far as arbitraging is concerned, several authors [for example Aldridge (2010) and Mitra et al. (2016)] suggest HFT bringing market inefficiencies, in so doing undermining the Efficient Market Hypothesis, but those studies provide little or no quantitative analysis of this phenomenon, therefore further research seems appropriate.
4. Similarly, the effect of Stop Loss orders has been pointed out by many scholarly and institutional studies [Cliff, Brown and Treleaven (2010), CFTC-SEC (2010a, 2010b), Foresight (2012)] but, again, no quantitative demonstration has been brought in before this research.
5. Although the topic of frequent order cancellation has been addressed by many scholars [Hasbrouck and Saar (2013), Brogaard, Hendershott and Riordan (2014), Menkveld and Zoican (2016)], the issue of price change due to cancelled liquidity is not dealt with in depth, and the role played by exchange speed versus traders’ speed is definitely a gap in the

literature. Indeed, it is mostly addressed by the exchanges themselves when boasting improvements in their own press releases.

Lastly, the combined result of two or more of these effects is rarely, if ever, explored. This research intends to address the least developed topics and the gaps found in the literature. The next section presents the purpose of this research and the following five sections illustrate each of the five topics dealt with in depth (and with different methodologies) in the following chapters.

3.1.2. PURPOSE OF THE RESEARCH

The high degree of uncertainty on many of the HFT-related topics summarised in the previous section leaves space to research. The purpose of this research is: (i) to produce an in-depth data analysis and computer-based simulations of the market environment to investigate whether, and if so how, financial stability is affected by the presence of fast investors aside the slower ones; (ii) to verify how HFT and financial stability interact with each other under non-linear conditions; (iii) whether, and if so to what extent, apparently innocent behaviours can lead to potentially destabilising effects, even in absence of illicit or illegal practices; (iv) to provide quantitative support to the theses, either from the audit trail data or resulting from simulations.

Computer-Based Trading is deemed to have the potential of creating a non-linear financial system [Foresight (2012)]. Accumulation of, and complex interaction between, different factors operating in the financial markets are likely to lead to a non-linear sum thereof, generating the well-known 'butterfly effect', as stated by the Theory of Chaos according to many authors, Gleick (2008) among them. It can be argued that at very high speed (that is, when control mechanisms have less chances to address potential issues), system sensitivity may be positively correlated to speed - and financial systems would prove to be no exception. According to CFTC-SEC (2010a), a convincing explanation of the Flash Crash should address several potential causes. This research intends to bring a contribution to a better understanding of the impact of HFT on financial stability. To this purpose, the next chapters present a few simulations showing the impact of HFT on some aspects relevant to market stability: the simulation presented in chapter 4 focusses on the impact HFT has on volatility, the one in chapter 5 investigates the possibility of HFT creating, under certain conditions, two market tiers, and chapter 6 discusses a simulation showing the effects HFT has on the Efficient Market Hypothesis. Chapter ‘7. Flash Crash Data Analysis’ carries out an analysis of audit trail data in support of the results obtained with the volatility simulation and an analysis of the impact Stop Loss orders might have had on the Flash Crash. Naïve orders, a concept never found in the literature, are also addressed in chapter 7, as an indication of market anomalies in presence of ultra-fast trading and as an indirect confirmation of the findings of chapter 4. Moreover, chapter 7 also illustrates the issue of relative versus absolute speed as an exacerbating factor of order cancellation, ghost liquidity and their effect on volatility.

3.1.3. IMPACT OF HFT ON PRICE VOLATILITY

As seen in chapter 2, some literature acknowledges an improvement of market stability parameters, and volatility in particular, as more HF traders enter the game, although this is by no means a unanimous view. A major concern against this view comes from the observation of the different latencies experienced by the fast traders and the traditional ones. The High-Frequency traders make use of the latest technology, including multi gigaflops computers, automatic news reading devices, lean software, Field Programmable Gate Arrays (FPGA) technology, ultra-fast or dedicated networks, and co-location. The net result is that the same physical object, the market, is operated upon by two categories of actors, each running at greatly different speed, in the order of magnitude of 1000:1 ratio or more. This means that even the simplest operation a fast player performs, may take a comparably long time when performed by a slow trader. Quote reading is an example of such a simple operation. Given the above-mentioned speed ratio, while a LF trader reads a price, another trader working at high-frequency may read 1,000 of them. Or, more likely, reading the prices and spending the rest of its large time advantage to act upon those prices, potentially leaving its slower counterpart with obsolete, and therefore useless, pricing information. But there is more. Once the read-and-act process in the LF trader's mind has been triggered, there is no logical reason for the trader to stop it before it terminates. Let us assume the price is deemed appealing at the time of reading and a LF trader starts the process of sending an order. Even if in the meantime the price changes the trader has no reason to stop it from going ahead, unless it decides to perform another read operation to update its pricing information. There is no guarantee that the price acted upon by a LF trader will still be valid when its order hits the book. So, the slow traders will always send an execution order for a security based upon potentially obsolete information. If the liquidity does not evaporate in the meantime, it will still be able to trade at the originally intended price. But if the liquidity vanishes, because the faster orders arriving before the one posted by the slower participant have consumed it, the latter's market order will execute 'at best', even if the 'best' price is worse than originally intended. Leaving alone the damage suffered by the LF traders (it could even be an advantage, if the fast orders moved the price in its favour), it is likely that the volatility experienced by that security unintendedly increases. The literature surprisingly shows little quantitative research taking this issue into account. In a simulation, the occurrence described above could be implemented by use of a queue, where Low-Frequency traders' order are temporarily parked until a suitable time period has expired, in order to simulate their latency. Then they shall be executed at the prevailing price at that time, with the consequence of moving the market in an unpredictable way. In the next chapter a simulation of this scenario shall be described in detail and its outcome analysed with the help of the appropriate statistical tools with the purpose to verify if the situation depicted above can actually occur or it is just mere theory.

3.1.4. HFT AND MARKET TIERING

The effect of different latencies experienced by HF and LF traders described in the previous section is that the two trading communities drive on different lanes. The members of the fast community enjoy a much more precise view of the prices at each instant in time whereas the members of the slow community carry out wishful guesses - and act upon their hopes in a sort of blind trading. Some authors, like Brogaard (2011) or Cartlidge and Cliff (2012), suggest this may lead to HF and LF traders only, or mostly, dealing within their respective communities. In order to understand whether or not this is a description of what really happens on the financial markets a computer simulation shall be implemented. The discriminant between the fast lane and the slow one is the bid-ask spread. HF traders are known to prefer tight spread situations for posting aggressive orders, and instead quoting mostly passive orders when the spread is wide. This can be explained with the higher probability of profitable aggressive trading when the thin spread makes it more likely to close the position as soon as favourable small price swing materialises. Conversely, a larger spread tends to favour passive orders taken up by investors buying at the ask and selling at the bid. A suitable number of either passive or aggressive orders is simulated and all the information about those trades are recorded onto a database. Each trading is assigned either to the thin spread or wide spread category, according to whether it was executed with a spread equal to one tick or more than one. The analysis verifies whether abnormal level of intra-community (i.e. HF-to-HF or LF-to-LF) trading occurs. In order to restrict the cause of whatever behaviour found in the simulation, another simulation will be launched, with no speed difference between the two communities. This way it will be possible to verify whether the results obtained with the previous approach were depending on the spread preferences or on relative speed. This simulation shall be described in detail in chapter 5.

3.1.5. IMPACT OF HFT ON THE EFFICIENT MARKET HYPOTHESIS

The Efficient Market Hypothesis (EMH) was established by Fama and Samuelson in the 1960s. According to the Hypothesis, prices encompass market information and it is therefore impossible to consistently make abnormal profits, above the ones achievable with a simple buy-and-hold strategy, unless taking more risk. Whenever an inefficiency in securities pricing arises, arbitraging will immediately sweep it out, re-aligning prices to their fundamental value. The EMH still holds in presence of arbitraging because the abnormal profits would not be consistent, being them shared by the very many market participants. However, a relatively new phenomenon, arisen in the last ten years or so, has put the EMH under severe questioning - that phenomenon is High-Frequency Trading. Although price discrepancies between markets, related instruments or related securities have always been observed and exploited by arbitrageurs, HFT allows a limited number of traders to beat the rest of market operators by exploiting their higher speed, due to superior technology, faster networks and co-location. If the abnormal profits caused by arbitrage are shared by a small number of High-Frequency traders, the assumption that nobody can systematically beat the market does no longer hold and doubts could be casted on the Efficient Market Hypothesis.

Chapter 6 presents a simulation made up of two markets sharing similar rules and trading the same securities, in which participate a small number of High-Frequency traders and a large number of slower ones. When in either market a price inconsistency arises a trader will swiftly sweep it out, consistent with the EMH. The purpose of the simulation is to verify whether HF traders do make consistent, abnormal, risk-free profit (usually at the expenses of the slower ones) and, in case such risk-free profit does exist, whether it can be dismissed as a random occurrence or is it statistically significant. Since Fama (1970) states that transaction costs would reaffirm the validity of the Hypothesis even in presence of arbitraging, different level of transaction costs are tested to appreciate the quantitative aspects of the simulation and its validity, or otherwise, in times of High-Frequency Trading.

3.1.6. DATA ANALYSIS OF THE FLASH CRASH

The investigation carried out in chapter ‘7. Flash Crash Data Analysis’ aims at verifying whether Stop Loss orders had a significant impact on the Flash Crash. A report by the Bank of Canada, Barker and Pomeranets (2011), identifies in cascade selling due to Stop Loss orders a possible cause of price movement exacerbation. Paragraph ‘7.2. A Simulation Using Petri Nets’ presents a model to show whether the theoretical possibility of such an occurrence exists. The following paragraph investigates the matter from an empirical point of view by analysing audit trail data. This is by no means a minor issue: cascade selling attracts suspicions of manipulative practices (not dealt with in this research) or systemic factors, potentially leading to repeated problems. The number of mini flash crashes identified by the literature (section '2.2.15. Flash Crash' Lesser Brothers') seem to point in this direction. The Flash Crash was but the most noticeable of a stream of events that, if not prevented, may jeopardise the confidence of the general public in the orderly functioning of financial markets. This is the reason for many financial institutions, regulatory bodies, research centres, governments and investigation bureaux paying so much attention to such disruptive events. And this is also the reason for the need of deep understanding of any, no matter how remote, possible cause of abnormal price movements. Another issue investigated is the one labelled 'naïve orders'. Chapter 4 argues that slow traders may post market orders at less than optimal price because of their intrinsic delay. Since the amount of details in the data does not allow to directly inspecting this phenomenon, similar naïve occurrences will be searched in limit orders (easier to spot out from the available data). It will then be assumed that if an abnormal number of naïve limit orders occurred under certain conditions, chances are that market orders were also affected by the same naivety, making it possible to compare simulation outcome and hard data.

On a more down-to-earth level, exchange trading servers regularly carry out a lot of operations for granting smooth market functioning. In order to ensure that the large amount of traffic is being dealt with within a reasonable time and in an orderly manner, several CPUs are used. So, it may well happen that while one CPU is dealing with a trade another CPU is updating the order book. Sequential operations would not be acceptable at today's fast market pace. In particular, while one CPU is updating the top row of the book with the outcome of a trade, another may be updating another row by cancelling an outstanding limit order. The combined effect would be a sudden and massive disappearance of liquidity. Therefore, chapter 7 also investigates whether or not these issues actually materialised during the Flash Crash. In order to do so, this research shall analyse the

Market Depth Data for the E-mini S&P 500 futures contracts, reporting the ten best bid and ask quotes, cancellations, and all the trades occurring between 14:39:00 and 14:45:28.115 of May 6, 2010 and compare them with the same amount of event occurred in the previous three days and in the following one.

3.1.7. RELATIVE VS. ABSOLUTE SPEED

The speed advantage discussed in the literature nearly always relates to a relative speed: if trader A is even marginally faster than trader B, it will (other things being equal) consistently overcome its competitor. It can said that speed is always relative. However, it is important to point out the difference between the relative speed between two traders and the speed of a trader with respect to the exchange it is operating at, that can be taken, by comparison, as absolute – a difference rarely dealt with by the existing literature. In one-exchange environment the speed of the exchange server can be taken as absolute, because it is given in that environment, and the speed of all other entities as relative to it. Insufficient speed by the trading engine with respect to the speed of the traders could cause abnormal behaviour in price movements. It may result insufficient for external reasons (e.g. high traffic) or for internal ones. This point is developed in Barker and Pomeranets (2011), who highlight that "[s]ystemic risk could arise if the market infrastructure fails to handle the large volume of transactions. [...] The increased strain on market infrastructure could lead to latency in both pricing and trade settlements" (ibid. p.51). It is true that the speed of the exchange server is not necessarily constant over time, as it can fluctuate because (among many other reasons) of the amount of traffic it handles, but it can still be considered absolute at any instant in time, even if it is different from an instant to another. The crucial distinction between the two kinds of speed has been pointed out by Farmer and Skouras (2012a): "the private benefits of speed come from relative speed (i.e. being faster than others) while market quality is determined by absolute speed levels" (ibid. p.3). A different argument still revolving about absolute speed is developed by Durden (2010), who sharply notices how on May 6, 2010 the NYSE started to trade at prices slightly below the NBBO.

The explanation provided is NYSE quote prices lagging behind those of other markets, with the result that it was (falsely) showing higher bid prices which attracted frantic sellers, just to realise the actual price was lower than the one they observed. This raises the issue of a different kind of speed: no longer the speed of traders relative to other traders but the absolute speed of exchange servers contributing to the NBBO. This is a topic that can occur whenever the amount of traffic is higher than the capability of the server to handle it. It is true that servers' technology is always improving, newer servers can replace older ones, and that more servers can be put in parallel operations to multiply processing power. Yet, simply adding more and more client computers could potentially lead to a critical point when the traffic will be heavier than no matter how many servers can cope with. Moreover, the computing power of the exchange servers can be considered pretty much static (in order to change, it needs long advance notice) while the number of clients of that server can increase dynamically at any time, with no notice at all. Anticipation of this issue was found a year before the Flash Crash by Budimir (2009) who argued that "an increase in trading activity implies disproportionate increases in Xetra [the eXchange Electronic TRAding system of the Frankfurt Stock Exchange] latency" (ibid. p.54). A similar point is also raised by Labuszewski (2010): “although telecommunications systems can be very fast and operate with very high capacities, systems may nonetheless be taxed with high message traffic in active markets.” (ibid. p.70). A standard behaviour of exchanges, as explained among other by Wah and Wellman (2013), "[w]hen a new order matches an existing order in the order book, the market clears immediately" (ibid. p.5). However, when discussing what happens in a HFT environment, and the word 'immediately' is used, the first thing to question should be its real meaning. What it means for an occasional day-trader equipped with a home computer may well be different from a major financial institution with professional CBT equipment, which on its turn is very different from a co-located HF trader posting and cancelling a few thousand orders per second. So, 'immediately' may not be so immediate for everyone after all, depending on whom this word applies to. When it applies to exchanges, a lot of care should be taken. In that case 'immediately' matching a market order with an outstanding limit order means that no other operation takes place before the trade is executed. But this is only true for the CPU carrying out that trade, and not necessarily for the other CPUs in charge of receiving, posting and cancelling orders. In particular, the Stop Loss mechanism is a source of exchange latency because it is handled just after a trade at a certain price occurs. An investor may quote a price and simultaneously specifying its will to close the position as soon as a certain loss level has been accumulated. This means that if a trade occurs which hits that Stop Loss price, the exchange must ensure the Stop Loss order is converted into a market order and executed ‘immediately’. While this sequence of operations occurs, no matter how fast, the rest of the world will hardly sit waiting for the exchange to complete them. Many things could happen in the meantime, and some of them could potentially affect the price the Stop Loss order executes at. So, the scenario may change under the feet and only being revealed when it is too late to act upon.

Another goal of this investigation is to understand whether HFT, coupled with insufficient absolute speed, may lead the market to behave erratically. In order to demonstrate if such a potentiality exists, irrespective of whether it actually occurred in a specific market or at a specific time, a suitable mathematical model representing that scenario will be illustrated.

3.2. A CHOICE: WHY USING SIMULATIONS

3.2.1. INTRODUCTION

Research in economics and finance is largely based on econometric models. Mathematical models are ubiquitous in academic publications on economics and their importance gave rise to econometrics as a discipline on its own right. Depending on the different disciplines it is applied to, academic research may directly observe the phenomena of interest or using a more indirect approach, for a variety of reasons. An example of the former approach is observations of planets, stars and galaxies. The objects of scientists' observations are the actual target of the research. In many other cases, despite being much closer to the observer, the objects can only be observed indirectly. Social sciences, economics, and finance in particular, very often work on relatively small samples because counting the very large number of objects the research focuses on may simply not be feasible. Researchers build models as realistic as possible and then observe their behaviour. This introduces a further degree of uncertainty: to the reliability of the observation it must be added the reliability of the model observed. As noticed by Reiss (2011), "[i]nferring from data to phenomena consists in disentangling what is a property of the specific setup used to produce the observation and what is a property of the phenomenon of interest" (ibid. p.251). Practitioners work on different definitions of what a simulation is. Some interpret simulations as using computers for solving equations that cannot be solved analytically. This is the task akin to numerical analysis and is not what will be pursued here. Another common use of simulations, which includes the one adopted by this research, is mimicking the behaviour of a real-life process by a computer process. A further interpretation, more general but more focussed on economics, is the one provided by Reiss (2011): "Simulations in economics explore the properties of computer-implemented models; they are aimed at drawing inferences about properties of a socioeconomic system [...] of interest" (ibid. p. 245).

In the following sections, some opinions in favour and contrary to using simulations will be discussed, with emphasis on authoritative advice on how to conduct them if robust results are to be achieved.

3.2.2. UNDERVALUED SIMULATIONS

Although economics and finance very often study phenomena by using an indirect approach, simulation is an undervalued technique. Reiss (2011) reports having conducted a search over the period 1969-2006 on EconLit, a database of economic publications that covers hundreds of journals, books, PhD dissertations, and working papers. The number of works containing the word 'simulation' (or derived from it) was above 900 per year in the last decade of the search period. Yet, this amounted to less than 3% of all the items in the database. A reason was provided by Lehtinen and Kuorikoski (2007), who argue that economists do not like simulation as its outcome will never be as irrefutable or as general as the outcome of a mathematical deduction, being the former a function of a specific combination of parameter values. It is also true that building appropriate simulations requires a great deal of previous knowledge about the phenomenon being studied. Having said that, Reiss (2011) identifies and refutes two common objections as resulting from biases.

(i) Simulations should only be used to test the plausibility of existing theories. The article claims that a simulation can do much more than just that, "by adding a variety of model building tools that help to expand and improve theory and make it applicable to concrete phenomena" (ibid. p. 254).

(ii) Simulations and real experiments differ in their materiality. Although some animals share a high percentage of genetic patrimony with humans, animal models are notoriously bad predictors of toxicity in humans. The article goes as far as arguing that the causal relationship between smoking and lung cancer was for many years in doubt since no trace of it was found in experiments on laboratory mice.

It may be argued that economics does not make much use of ‘real experiments’, being it more focussed on econometric models or data; however, the criticisms mentioned above still apply to simulations as opposed to mathematical models or direct observation of real-world data.

3.2.3. STRENGTHS OF EMPLOYING SIMULATIONS

Despite being simulations, by and large, disregarded by the economics and finance academic community, they display several characteristics that should make them desirable. Indeed, simulations have many advantages even over paper-and-pencil mathematical models, to the goal of better understanding the behaviour of a socio-economic system. As stated above, some economic phenomena involve such a large number of actors that direct experimental access becomes infeasible, and even small economic events may be hard to experiment about for practical, technological or ethical reasons, not to mention costs of the experiment. Reiss, a philosopher of science, is interested in the conceptual aspects of the matter, and finds in the less stringent idealization requirements an evident advantage of simulations vis-à-vis mathematical models as well as laboratory experiments. In his opinion, a simulation is much more flexible than the corresponding mathematical model since the former requires fewer assumptions to produce results. Therefore, simulations can be used to successfully address the problem of over-constraining assumptions, another issue strongly felt by philosophers of science.

The advantages listed by Reiss (2011) also include the possibility of precise replication, as the same computer program fed with the same input data will deliver precisely the same result, differently from laboratory experiment or real-world data, where controlling the environment is not always as easy. Moreover, a simulation can easily vary some parameters that cannot be varied in nature, or even assigning them unrealistic values, for sake of checking boundary (or out-of-boundary) conditions. Lastly, usually costs are lower and implementation speed is higher. Keller (2003) notices how several years of practice have shown that computer simulations can not only explain 'how' things happen, but in many case also 'why' (the ultimate frontier of knowledge). The ambition confessed by Reiss is that "computer simulations can help economics on its way to a full-fledged experimental science" [Reiss (2011), p. 244].

3.2.4. WEAK EXPERIMENTS, ROBUST SIMULATIONS

Scientific experiments and real data observations have been conducted for the whole history of science and scientists are well aware of their limitations, including the interference a measurement instrument can cause to the tested environment. Walls in a wind tunnel introduce different kinds of turbulence from those experienced by cars or aircrafts under operating conditions. Similarly, human traders are not under the same operating conditions (that is, not so much under stress) when interviewed by a researcher as they are when betting their or other’s money during a financial crisis, and their responses could be different in the two scenarios. These considerations may suggest real data observation as a better investigation tool; yet real data is given and immutable, and as such it lacks the flexibility of a computer-based simulation, for example in investigating extreme conditions. The differences between experimental and real environments can be curbed down to an extent by enhanced design and proper calibration but not completely eliminated. "[A]lthough a wind tunnel is in a sense 'more similar' to the target than a simulation model, we have no guarantee that inference conclusions drawn from wind-tunnel experiments will be more reliable than conclusions reached on the basis of simulations" (ibid. p. 255). Nigel Gilbert, a great supporter of the use of simulations in social sciences, was the first to propose a 'policy wind tunnel' for testing in a simulated environment the impact regulator's decisions might have on financial outcome in the real world. In order to ensure that the technique is properly adopted and robustly conducted, he suggests a few further enhancements [Gilbert (2008)].

(i) Adding assertions: Checking that the outcome of the simulation falls within acceptable range of values by adding as many assertions as reasonable to establish that range.

(ii) Testing with parameter values for known scenarios: checking the outcome against sets of known input-output values (often available) to ensure reliability and enhance trustworthiness also for the unknown scenarios.

(iii) Using corner testing: checking the outcome for boundary values, those that represent 'black swans' and stretch the simulation to extremes.

3.2.5. CONCLUSION

Generally speaking, simulations are not widely adopted in economics since they lack mathematical elegance, analytical approach, and certainty of results. Definitive advantages econometric models have over computer simulations are rigour and reliability. On the other side, these plusses are paid for in terms of reduced scope, lesser flexibility and sometimes higher monetary costs with respect to simulations. The so-called 'aggregation problem' is another source of concern to mathematical models. These models replicate the behaviour of an individual and her specific situation. But there is no guarantee that scaling both individual and situation up to macro level will properly depict the behaviour of a multitude of individuals, each with her own specific situation. There is no guarantee either that the composition of several linear phenomena will generate a linear phenomenon. Quite the contrary, the Theory of Chaos [Gleick (2008)] argues that in many cases the outcome is a non-linear behaviour. "The analogue for physics would be to model the behavior of gases at the macrophysical level, not as derived from the aggregation of molecules of randomly distributed momenta, but as a single molecule scaled up to observable volume" [Reiss (2011), p. 259]. What simulations certainly do well is modelling agents' heterogeneity. After taking into account the pros and cons, it could be said that simulation is neither absolutely better nor absolutely worse than an econometric model; most depends on the case under study and in many scenarios both approaches can live together and supporting, or contrasting, each other to the goal of reaching a more thorough understanding of the socio-economic or financial phenomena.

The purpose of this research is to bring a contribution (not necessarily the definitive, but a useful, one) using this less-explored path, to a more comprehensive understanding of the impact of HFT on financial stability. It must also be noticed that most of chapter ‘7. Flash Crash Data Analysis’ is dedicated to real world data analysis.

3.3. A CHOICE: SELECTION OF THE ENVIRONMENT

A simulation, as any other investigation, has first to define its field of application and its scope of validity. Simulating the Flash Crash could involve replicating the entire US securities market: but that is clearly out of the scope of this thesis. Too many different exchanges, with too many different rules and, above all, too may CPU behaviours to mimic. None of the simulations discussed above took that approach. A feasible choice was to restrict the simulation to one market but if a simulation had to consider only one market, it had to be a very significant one – the one where everything started and everything ended; the one which entailed the *Alpha* and the *Omega* of the Flash Crash. After some analysis of the events on the day of the Flash Crash, the environment selected for modelling was the E-mini S&P 500 futures June 2010 contracts on the Chicago Mercantile Exchange (CME). It is true that on that day, many exchanges in the US showed erratic behaviours but according to the CFTC-SEC (2010a, 2010b), everything started off in Chicago. It was there that the investment firm Waddell & Reed started to sell 75,000 E-mini S&P 500 futures contracts, for a total value around $4.1 billion, via an algorithm offering contracts worth 9% of the amount traded during the previous minute but regardless to price or time. As we have seen before, the CFTF-SEC conclusions have been severely criticised but even so, it is undeniable that the beginning of the recovery cycle (albeit with a lot of ups and downs), started when the 5-second Stop Logic mechanism was triggered by CME Globex platform. Even though the most striking effects occurred on other markets (like stub quotes hit as low as $0.01 or as high as $99,999.99), the CME still seems to have been in command of what happened on that day. CFTF-SEC (2010a) highlights the centrality of the CME by showing the sequence of E-mini S&P 500 Index prices split by hour, by minute and by second (no less than 14 pages of the report are dedicated to reproducing that information, and no other market has been dedicated so much space). Bloomberg, cited in CFTC-SEC (2010a), published a graphic that became the visual representation of the event. It plotted the price of Dow Jones Industrial Average, S&P 500, and E-mini S&P 500, versus time of the day, as in figure 1. Also Kirilenko et al. (2011) and Menkveld and Yueshen (2016) assign a central role to the E-mini S&P 500 futures market in the crisis. The choice of the CME for this research turned out to be an appropriate one since, despite the fact that oscillations in the S&P 500 stocks would reverberate in the related futures traded at CME, the E-mini futures are not traded on any other market, something which limits, as far as practically realisable, interaction and price disturbances from other venues. As a second milestone, the decision to restrict analysis to the six-and-a-half minutes between 14:39:00 and 14:45:28 of May 6, 2010 was taken. As shown in figure 2, three points in time were the natural candidates for starting the analysis: 14:39, 14:40 and 14:42. After some thought (and some tries) the decision was to go for the first one. This choice involves more data to handle (more than 580,000 market events, including quotes and trades) but it reduced the risk of important events falling outside the interval of investigation. The end of the interval was easier to choose: the triggering time of the CME Globex Stop Logic at 14:45:28.115 was a natural choice. For ease of referencing, all the previous times refer to the time in New York (Eastern Daylight Time - EDT), whereas the commercial data provided by the CME are all timestamped with reference to the Greenwich Mean Time (GMT).

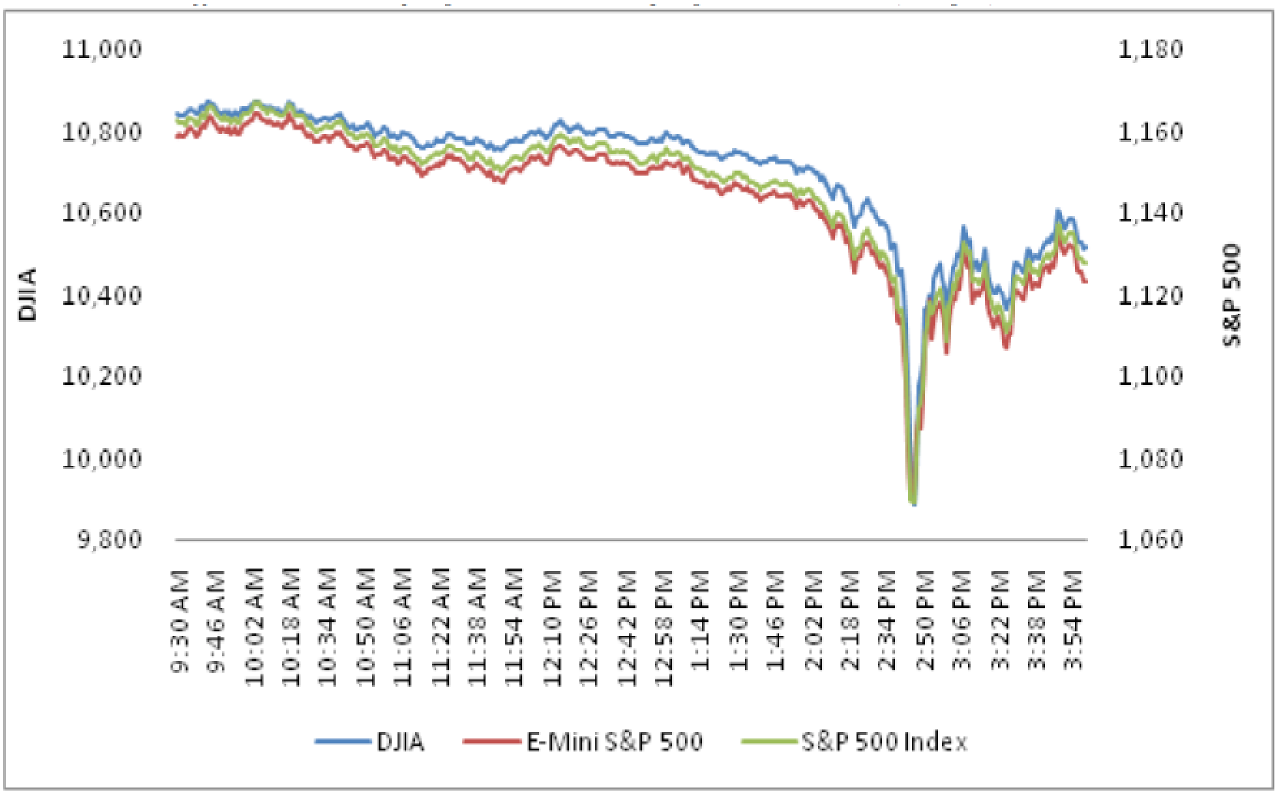


Figure 1. Equity Indices and Equity Index Futures, May 6, 2010

A detailed explanation of the main rules governing the trading at Chicago Mercantile Exchange and a schema representing its architecture can be found in Appendix D.

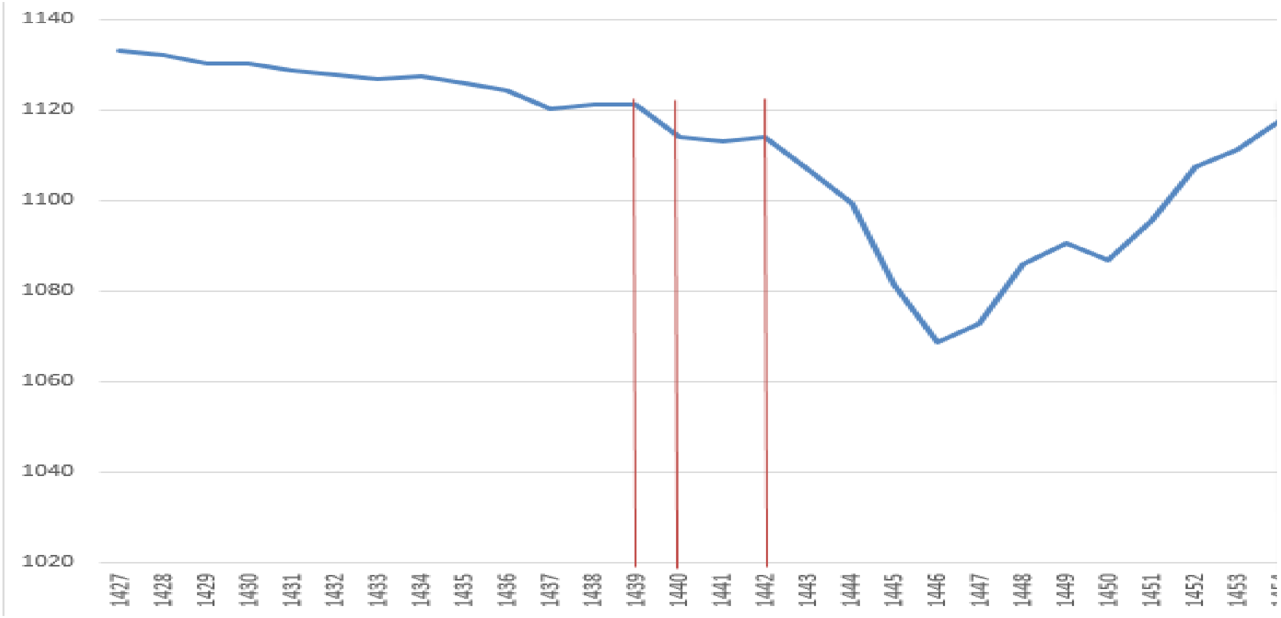


Figure 2. Price of the E-mini S&P 500 futures contract at the CME on May 6, 2010

3.4. A CHOICE: PARALLEL VERSUS SEQUENTIAL SYSTEM

One of the most important decisions that needs to be made is the parallel versus sequential simulation of the market. A market environment, in its simplest (and only conceptual) form, typically consists of several market operators placing limit or market orders, a trading engine, and two databases, representing the limit order book and the trading records. In real life, the booking-trading process is definitely a parallel one. Several traders post orders, or match posted orders, on several securities, potentially at the same time. Orders are usually dealt with according to time priority: first come, first served, even though in some cases there are slightly different arrangements, like proportional filling of limit orders against an incoming market order. When two different orders reach the database at the same time, then it is up to the database management system (DBMS) to make a decision between the two and to act in a way that is regarded as fair by the operators and the regulatory authorities alike. A quick digression about the expression 'at the same time' is now needed. Two events happen 'at the same time' when the time granularity of the system’s clock is larger than the time difference between the two events. If my source of information is the daily newspaper, all yesterday's events happened, in my view, 'at the same time', because I am not able to assign them any chronological order within the time window - in this case, a day. If I watch at a Bloomberg screen, 'the same time' is restricted to all events occurring between two consecutive screen refreshes. In computer science, this is relatively straightforward as the database input mechanism works in a sequential manner, an example being a receiver agent listening on each of the input ports in turn (as depicted in figure 3). So, if there are N input ports, the database input engine scans each of them in a sort of circular queue and if two ports receive data from the external world to be written onto the database at the same time, the first one on the queue shall be served. The sequential fashion of this process may not be appreciable by a human operator because of its extremely fast pace, but true parallelism is still an illusion. Yet, although a sequential model would satisfactorily replicate a parallel system, in order to more closely simulate the behaviour of the market, a parallel approach would be advisable. However, this approach would present two kinds of problems. First, it would be rather difficult to analyse the behaviour of the system and, second, it would require several operating system processes interacting to each other via standard mechanisms like inter-process communication or similar.

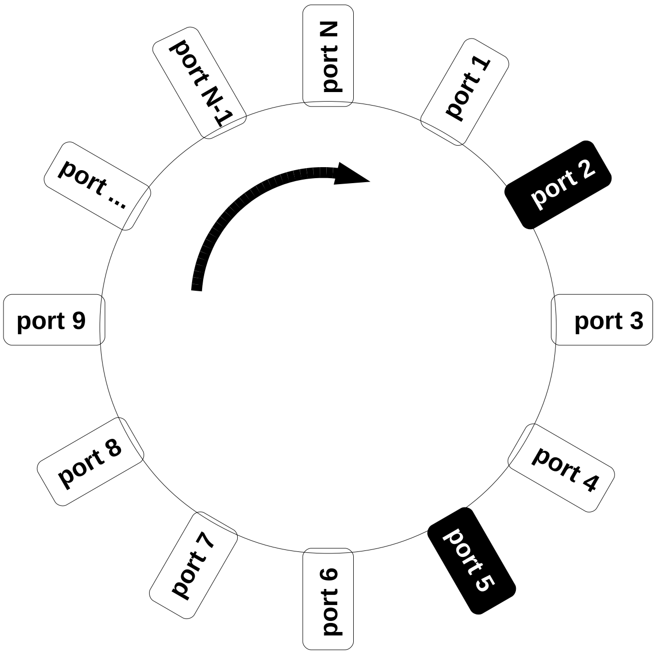


Figure 3. As the control (depicted as the arrow) is now about to check the content of port 1, data on

port 2 will be picked up before data waiting on port 5, even if data on port 5 arrived before data on port 2.

This way, even if in practice the CPU treats data sequentially, the visible effect at the macro level is parallelism. A different approach is using an intrinsically sequential approach, where only one operating system process is involved, with a closer resemblance of the micro level behaviour, where understanding how the system works would be much simpler. One may argue that multiple CPUs can, and actually are, used in a real automated trading process. Yet, if exclusive access to a datum is to be guaranteed, than no matter how many CPUs are insisting on it, a sequential mechanism has to be devised at some stage. As a clarification we may think of an airline seat booking system. Several travel agencies may access the booking system for the same seat but, since that seat is unique, at some stage sequentialisation must occur, in order to avoid double-booking of the same seat and disappointment of all customers (who would fail to sit the place they were guaranteed to have booked) except a lucky one. Similarly, if a security were on offer and several potential buyers posted market orders for it, one and only one must be served, or the market integrity would be in jeopardy. It is widely accepted that market operators must be granted the maximum possible freedom to act on the books but at least the database management system must ensure orderly access to data in write mode, that is, for changing the status of the books. Then, after having considered the trading system simulation, also the difficulties of the post-mortem analysis must be evaluated. It is out of doubt that analysing the output of a parallel system is harder than the output of a sequential system. In the latter case, one is always sure of what is going on, who is acting, what is doing, when, and so on, whereas in the former case the decidability may be more complex. If two or more operators are acting at the same time the originator of a specific action may not always be clear, and multi-step actions (pretty common in real life) even if logged, may be interleaved between two actors, making the work-flow complicate to follow and understand. A possible solution is to add a fine timestamp and an actor's ID to each and every action logged but this would require an extra step of re-constructing the original workflow for each of the actors in the process, and moreover two events may still occur within the same time unit, however small the granularity of the system (millisecond, microsecond or even nanosecond), making it difficult to discriminate between the originator of each event. In summary, on one side, a parallel system may more closely simulate the macro behaviour of the markets whereas a sequential approach is simpler to implement, in general without loss of functionality. The latter was indeed the preferred approach.

3.5. A CHOICE: MODELLING A SECURITIES EXCHANGE SYSTEM

3.5.1. DEFINITION OF PETRI NETS

Another important choice to make was the selection of a suitable framework for modelling a generic securities exchange system, used in paragraph '7.2. A Simulation Using Petri Nets' and in section '7.4.4. Impact of Exchange Latency Using Petri Nets'. Petri Nets is a powerful tool for modelling complex, dynamic, concurrent, asynchronous, distributed, parallel, non-deterministic, stochastic systems in a formal way [Murata (1989)]. The original concept is due to Carl Adam Petri (1926-2010), a German computer scientist who created this technique that rapidly won widespread acceptance among the academic world and the practitioners' community alike. Since the time it was first introduced, many variations of the Petri Nets have seen the light of the day and in the literature there are a few different formal definitions of what a Petri Net is. Although the details are varied, they all share the same underlying idea which, in what could be one of its simplest formalisations, can be summarised as follows: A Petri Net, N, is a 3-tuple (P, t, f), where P is a finite set of *places*, t is a finite set of *transitions*, and f represents a *flow relation* (P x t) <union> (t x P) -> N. The extra condition, P <intersection> t = <void>, imposes that the sets P and t are disjoint. A Petri Net works by taking an initial number of tokens in some places (called initial marking) and moving them to other places via the transitions. Places and transitions are connected via directional entities called *arcs*, representing the flow relation. A transition can fire only if all its input places are marked with (at least) one token and its effect is removing one token from each of its input places and adding one token to each of its output places. The set of input and output places linked to a transition is given by all places connected via its incoming and outgoing arcs, respectively. A pictorial example of what 'firing' means is given in figure 4.

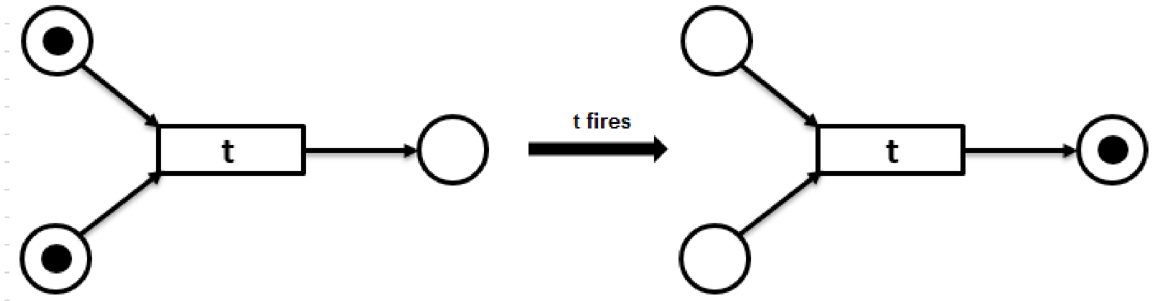


Figure 4. Petri Nets firing mechanism

Many other definitions of Petri Nets exist; for example adding an initial marking to the tuple, or introducing the concept of weight function, which indicates the minimum number of input tokens required for a transition to fire. Other useful extensions of the Petri Nets are the Timed Petri Nets, which assign a timing to some transitions, forcing them to waiting a certain time to expire before firing (still of course with the constraint that all input places must have a token), or the Coloured Petri Nets, which add a qualitative flavour to the tokens themselves. For further details on the Petri Nets there is an extensive literature available, among others Murata (1989), Popova-Zeugmann (2013), and Reisig (2013).

3.5.2. PETRI NETS: A MATHEMATICAL MODELLING TOOL

A useful feature of Petri Nets is the ease to depict them graphically, where places are shown as circles, with tokens as smaller black circles inside places, transitions as bars or boxes, and arcs as arrows. Petri Nets are a mathematical modelling tool, in as much there exist a direct and straightforward way to convert a net into matrix form. The matrix dimensions are given by the number of places and transitions (where row M represents the *place* P_M and column N represents the *transition* t_N). In each [M, N] position, a positive integer *k* represents addition of k tokens by transition N to place M, whereas -*k* means removal of k tokens and zeroes stay for no arcs between the two entities. In the Petri Nets representation of the classical Producer-Consumer algorithm the

Producer, P1, owns a token, meaning it has produced and the Consumer, P2, is ready to consume as soon as the buffer, represented by place P3 will hold at least a token, representing the product. The matrix representing the Producer-Consumer net (shown in figure 5) is:

$$\begin{array}{ccc}
 & & \text{T R A N S I T I O N S} \\
\text{P} & & \begin{array}{cccc}\mathbf{t_1} & \mathbf{t_2} & \mathbf{t_3} & \mathbf{t_4}\end{array} \\
\begin{array}{c}\text{L}\\ \text{A}\\ \text{C}\\ \text{E}\\ \text{S}\end{array} & \begin{array}{c}\mathbf{P_1}\\ \mathbf{P_2}\\ \mathbf{P_3}\\ \mathbf{P_4}\\ \mathbf{P_5}\end{array} & \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}
\end{array}$$

Transition t1 picks up a token from place P2 (t1,P2 = -1) and adds one to place P1 (t1,P1 = +1), whereas transition t3 picks up tokens from places P3 and P4 (t3,P3 = -1; t3,P4 = -1), and adds one to place P5 (t3,P5 = +1). Notice that transition t3 can only fire if both places P3 and P4 hold a token.

The matrix properties of Petri Nets can be used to show the system behaviour. For example, in the situation depicted in figure 5 the only available change can occurs when transition t2 fires, that is, when the matrix above is multiplied by the vector [0, 1, 0, 0], where 1's represent transitions that fire. The result of the multiplication is a 5-element vector (-1, 1, 1, 0, 0) representing the changes occurred to the places. If we add this vector to the one representing the initial place situation, the sum yields the place situation after firing *transition* t_2:

$$(1, 0, 0, 1, 0) + (-1, 1, 1, 0, 0) = (0, 1, 1, 1, 0).$$

At this point either transition t_1 and t_3 can fire, which means that either vector (1, 0, 0, 0) or vector (0, 0, 1, 0) can be multiplied by the matrix representing the system, yielding the vectors (1, -1, 0, 0,

0) or (0, 0, -1, -1, 1) respectively. Adding to the vector resulted by firing transition t_2, we obtain either vector (1, 0, 1, 1, 0), after firing t_1, or (0, 1, 0, 0, 1), if transition t_3 was fired.

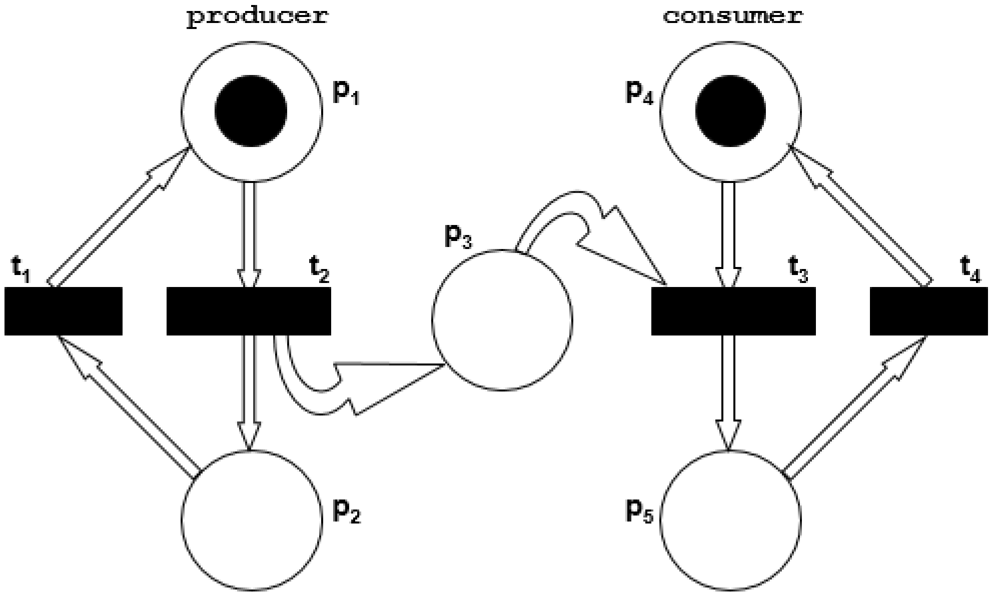


Figure 5. Petri Net representing a Produce-Consumer system

Repeating this process, the continuous functioning of the system can be reproduced and all the changes in the system can find algebraic description by matrix multiplications and vector additions. The matrix representation of the Petri Nets is useful as well for proving structural properties like liveness and boundedness of places.

An explicative example of the tool's power can be shown by using a Petri Nets implementing the mutual exclusion algorithm. The main drawback of the Petri Nets technique is its tendency to grow rapidly in size even for relatively simple systems. The Crosstalk algorithm is well known in computer science, being extensively used in undergraduate courses to show an instance of the mutual exclusion principle for concurrent systems (figure 6).

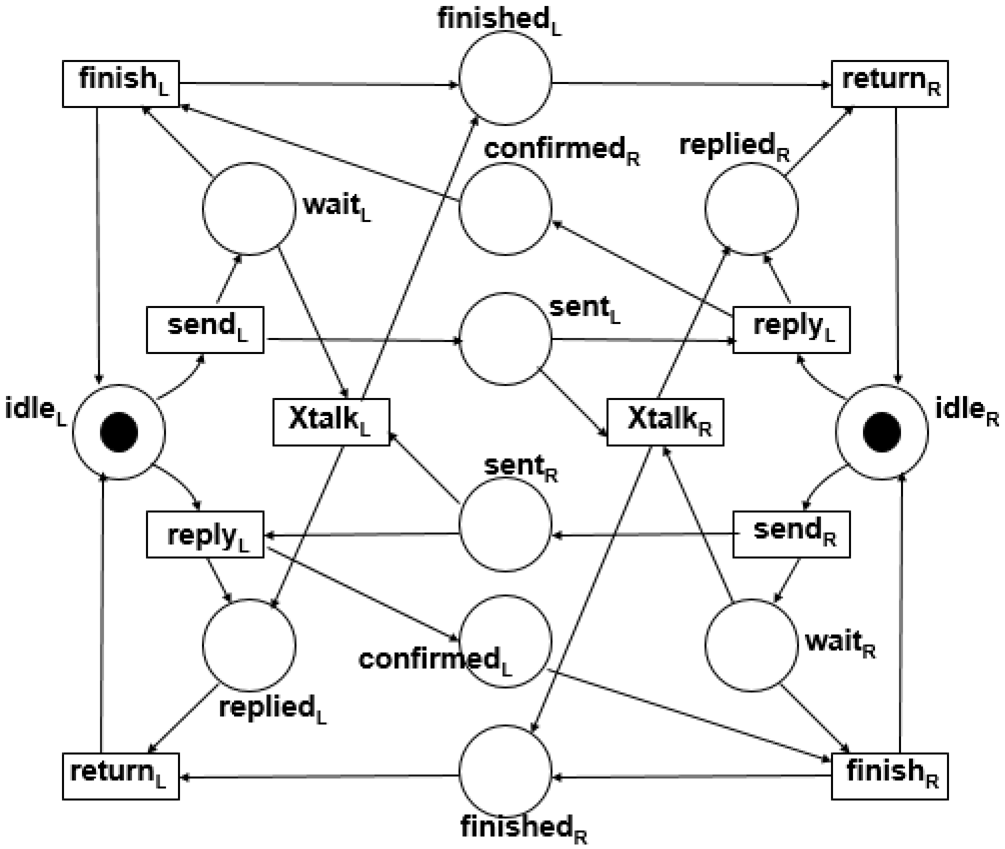


Figure 6. Petri Net representing the Cross-talk algorithm

The idea is that when one network component transmits its data, the other components in the same network cannot transmit their own until they have sent a reply to the message received by their counterpart. This apparently simple algorithm has a Petri Nets representation far from trivial. On the other side, one of the most appreciated features of Petri Nets is the ease of matrix representation and related matrix-based calculations. Petri Nets, originally thought as an aid to model communication in computer systems, have been successfully adopted over time in as diverse fields as data analysis, process control, workflow management, business science, biology, chemistry, engineering, securities trading, and very likely others.

In chapter 7 a model of a securities exchange system based on the Petri Nets formalism shall be used to demonstrate the interaction of dynamic, concurrent events occurring in an exchange server on a day-to-day (or millisec-to-millisec) basis. The purpose is to model an exchange system made up of bid and ask books, with and without Stop Loss features, a matching engine and an execution system. Moreover, each book must have its own cancel functionality for both plain orders and Stop Loss orders. The model, still a somehow simplified one, consists of 31 places and 26 transitions, most of them duplicated on the bid and ask side.

3.6. CONCLUSION

This chapter focussed on the gaps found in the current literature, on how they have been addressed in this research and on the operating aspects of the research; along with a detailed description of how it will be carried out in the next chapters, a few choices have been made and the rationale provided. In particular, the environment selected is the market for the E-mini Standard & Poor's 500 futures contracts traded on the Chicago Mercantile Exchange. This is a sensible choice because that market played a pivotal role in the Flash Crash and it does not suffer from direct interferences from other venues since the securities identified are only traded on the CME. Obviously, when a major crisis occurs, most markets and most securities are affected but the choice of the E-mini S&P 500 futures minimises the interconnectedness issues. The choice of a sequential, rather than parallel, system seems acceptable, although the implementation of a fully-fledged multi-venue, multi-CPU system is certainly a recommended path for future research. Lastly, Petri Nets represent a tool widely used for simulating dynamic systems such as stock exchanges. Chapter 7 will demonstrate how that tool can prove useful for highlighting market behaviours that real data analysis cannot grasp as clearly.

4. IMPACT OF HIGH-FREQUENCY TRADING ON VOLATILITY

4.1. INTRODUCTION

As discussed in previous chapters, one of the main aspects concerning financial stability is asset price volatility. The effect of HFT on financial markets, and volatility in particular, is controversial: as seen in the literature review (chapter 2), many renowned authors strongly support the view of HFT being a mitigating volatility factor but this is by no means a unanimous opinion. Overall, it can be claimed that the actual effect of HFT on market stability is still under investigation. This chapter has the purpose to bring a contribution for a deeper understanding of the matter and to concentrate on one aspect of market stability, price volatility, which is perceived as a serious risk factor, as at its extremes it may lead to potentially catastrophic events, like the crashes mentioned in previous chapters. The next paragraph describes the methodology used to simulate a market and verifying the impact HFT has on volatility. Then the content of the simulation will be described (paragraph 4.3), and its results presented and discussed (paragraph 4.4). Paragraph 4.5 concludes this chapter.

4.2. METHODOLOGY

As seen in previous chapters, academics have very different opinions about the impact HFT has on volatility. However, most studies approach the issue either from the purely theoretical viewpoint or by comparing volatility data for long periods, before and after entrance of HFT into the game. This research, as a few other studies did, bridges the gap between the theoretical and the empirical approach by simulating a market environment which experiences various levels of volatility according to different case and scenarios, and then verifying its results against the hard data in chapter 7. This is by no means the only methodology that can be used to investigate how HFT impacts on volatility: however, it is a method which carries several advantages (e.g. idealised environment, easy change of initial and boundary conditions, possibility to enforce specific behaviours) whose results shall be compared, and contrasted, to the results of other studies, so adding to the already existing amount of knowledge on this topic. The original approach used here is to acknowledge and explicitly implement in the simulation the delay low-frequency traders experience with respect to their HF counterparts, something not seen in other academic works so far. The Agent-Based Model (ABM) presented here has been built with the purpose to be as realistic as possible, and various features have been included with this goal in mind. The simulation is implemented with the aim to identify whether abnormal market behaviours can be detected, as far as volatility only is concerned, under two specific cases: a base case, with no particular disturbances in the market, and a trend case, when the market prices move systematically either up or down. Both cases use an Agent-Based Modelling approach to mimic the behaviour of the market under specific case. ABM is defined in Castiglione (2006) and the content of the webpage has been personally confirmed by the author in August 2015 as consistent with the current state-of-the-art. The base or trend cases are determined by the criteria used to set the three operating parameters: book type (bid or ask), order type (limit or market), and trader type (High-Frequency or Low-Frequency), according to the specific case. The difference between traders is mostly given by their relative speed. Whereas HF orders get serviced immediately, LF orders are subject to a delay, simulating their latency. As an example let us consider a sequence of order as follows:

1) HF limit order quoting price X. This order changes the book.

2) LF market order trading at price X. This order is queued up for later execution.

...) other orders being either executed immediately or queued up, according to whether submitted by HF or LF traders, respectively. Let us suppose that, because of the HF orders, the price of the security move to X+Δx

t) Eventually the order queued at step 2 gets served. The market order is executed at the prevailing price, which is now X+Δx. If the order was a sell, it enjoys a more favourable (higher) selling price than originally intended, while if it was a buy the trade is less favourable. In either case the volatility for that security increased without the explicit consent of the trader to deal at the new price.

Since the purpose of this chapter is to evaluate the impact of HFT on volatility, the simulation has been run under two scenarios: a market with only Low-Frequency traders and then a market with both HF and traditional traders. The results from the latter simulation are then compared with the scenario in which no HF traders take part, in order to verify the impact that HFT has on the market behaviour. All runs of the simulation follow the same standard structure but in each case the selection of the three main parameters, book type, order type, and trader type is made according to the features that are peculiar to the specific case implemented. The testable hypothesis for this simulation is whether the participation of HF traders increases the market volatility, and if so, under which conditions.

4.2.1. SIMPLIFYING ASSUMPTIONS

4.2.1.1. Strategy

No trading strategies are implemented, the only exception being the lesser probability of HF traders to post market orders when the spread is wider than one tick. This matches both Brogaard (2010), finding HF traders preference to trade on lower spreads, and the commonly-held view on HFT exploiting speed of trading while at the same time minimising the risks. The most well-known HFT feature is the tendency to quickly closing any open position, possibly with a profit, however small. In order to do so, HF traders exploit the common swings shown by the market but in order to take advantage of such minimal movements, the trading price must be very carefully chosen. If the bid-ask prices are, say, 100-101 and a HF trader buys at 101, chances are higher that a future upward movement will allow it to sell at 102 rather than at 103, necessary to make a profit if the initial spread was 100-102, instead. HF traders are still allowed to cross a wider spread (this occurrence is not prohibited in the software) but with lower probability that when the spread is thinner, and the lower the probability the wider the spread is.

4.2.1.2. Liquidity

The main simplifying features are about liquidity: (i) every limit order adds one lot of liquidity to the book and every market order consumes exactly one lot, and (ii) each price level has a maximum liquidity it can accept. The latter feature is useful for simulating the decision to quote an order at the next available price level. The simulation implements this feature, representing the expected behaviour of a rational investor, to mimic the following behaviour. When one price level has reached a certain amount of liquidity, set as the maximum allowed, any further limit order at that price is converted into a limit order at the next price level. The rationale behind this behaviour is that when there are many orders at a certain price level, this mechanism simulates an investor feeling reasonably confident of its and other investors' judgement to dare jumping to the next price level. For example, if the bid is at X and the liquidity is at its maximum, the next bid quote at X will be transformed into a bid at X + t, where t is the minimum tick allowed by the exchange rules. If the minimum ask quote is at X + t, then the limit bid order may or may not cross the spread and in the former case it could be automatically converted into a market order executing at that price. However, in order to make the simulation as realistic as possible, the implementation decision was not to allow spread crossing in case of limit order at the best price when liquidity is already at its maximum. In this case the incoming limit order is simply disregarded.

4.2.1.3. HF-to-LF trader ratio

Another common feature to all cases is the ratio between HF and LF participants. The ratio according to which high frequency traders are selected against their slower counterparts and their respective absolute numbers are read at initialisation time and remain fixed throughout the execution of the simulation. The HF-to-LF trading ratio is taken from the literature [among others, CFTC-SEC (2010a), Friederich and Payne (2011)], that report for a technologically advanced market, like the US, it to vary a lot but to be often around 50% of all trades. The simulation implemented does not allow to set the trading ratio directly, as it depends, in a random fashion, on the three parameters (book type, order type and trader type). Thus, a few attempts needed to be made, varying the ratio between the quotes posted by HF and LF participants (called quoting ratio, or QR) each time, until an acceptable trading ratio was obtained. The numbers of HF and non-HF traders are taken from CFTC-SEC (2010b), where at page 29 it displays summary statistics whence it is possible to extract the number of HF and non-HF (in the Report split between Intermediary, Buyer, Seller, Opportunistic and Noise) traders active in the E-mini contracts market on May 6, 2010 as well on the previous three days. The figures from the days before the 6th seem more sensible in order to avoid any reference to an exceptional situation as the day of the Flash Crash, namely 15 HF traders and 11859 slow traders.

4.3. DESCRIPTION OF THE SIMULATION

4.3.1. CASES AND SCENARIOS

The simulation represents a market in which traders can quote limit order or post market orders, either on the bid or the ask book. It is split into two scenarios: a slow market, with only LF traders taking part, and a mixed one, with both HF and LF traders. The former is useful as a comparison, as it represents the market 'as it used to be' before the appearance of HFT. The latter scenario shall be analysed to highlight the differences in volatility. In order to make the results easy to explain, it is paramount to restrict the differences between the two scenarios to the bare minimum; since they only differ because of the presence of HF traders, any statistically significant difference in the results can safely be attributed to them. The relative speed between HF and LF traders is implemented with a queue. It means that whereas a HF limit order goes straight to the book, and a market order gets immediately executed, order from LF traders are queued up for a pre-defined period of time for later execution, to simulate the latency that characterises that category of traders. During the time LF orders stay in the queue, the market carries out its activities as usual, adding new quotes or executing trades, if coming from HF firms or when the queueing period of a LF participant has expired. In the simulation two different cases were tested: the unperturbed (base) case and the trend case. The 'trend' case identifies a situation in which most market orders go to the same direction (either buy or sell) whereas in the 'base' case trade direction is random. Table 1 depicts the various scenarios and cases investigated, where QR indicates the quoting ratio between HF and LF traders (more on this later). A few simplifying assumptions needed to be made. These assumptions allow to isolate specific features to be investigated while still ensuring the overall behaviour of the simulation to represent a realistic market flow.

4.3.2. HOW THE SIMULATION WORKS

The simulation is launched 100 times, in order to provide robust results. Each of the 100 repetitions goes on for 10,000 cycles, each representing one activity on the market (quoting a limit order or posting a market order). The selection of the book upon which the next incoming order will act is randomly selected. The order type (limit or market) is computed randomly but, if the spread is wider than one tick, a higher probability is assigned to limit orders, in proportion to the spread.

		SCENARIOS	
		Low-frequency traders only	Both HF and LF traders
CASES	Base case (quiet market)	QR=0	QR=1, QR=2, QR=20, QR=200
	Trend case (market subject to a price trend)	QR=0	QR=1, QR=2, QR=20, QR=200

Table 1. Scenarios and cases investigated in this research

During a trend, if the randomly selected book is different from the trend direction (set at the beginning of that run), then only a limit order is possible in that direction, because of the assumption that during a trend no market orders in the opposite direction are allowed. The only restriction in place outside a trend is that no two consecutive market orders can be posted in opposite directions to each other: at least a limit order needs to be quoted in-between. This is to avoid a frenzy market with buy and sell trade orders with little underlying logic, which is against common experience. If the spread is thin, the trader is randomly selected according to the ratio established by the QR parameter read at the beginning of the simulation launch. If the spread is wider than one tick and a market order has been selected, then low-frequency traders are given higher priority than their HF counterparts, highlighting how HF traders feel more comfortable in tight markets. At the end of the 10,000 cycles, the difference between the highest and the lowest traded prices, an indicator of volatility, is recorded onto a database. Then, the average and the standard deviation are worked out over the number of launches. HF-to-LF quoting ratio equal to 0 (for the no HF traders scenario), 1, 2, 20 and 200 were tried, with the goal of finding a resulting trading ratio between 50% and 60% as suggested by the literature, as detailed above. In both the base case and the trend case, the 1-to-1 quoting ratio (QR=1) yielded an acceptable HF-to-LF trading ratio but the other ratios were also useful to verify the change in volatility as this ratio varies. It must be noticed that a QR=1 means that for every HF order there is, on average, also one LF order. This sounds sensible when taking into account that the HF/LF ratio is near to 400 [data taken from CFTC-SEC (2010a)]. If the order selected is a limit order and the liquidity is not at its maximum, the order is quoted at the originally intended price. In case the price is deep in the book, it is added to the existing liquidity. If the quote is above the current best bid or below the current best ask price, it is inserted at that price. If instead it was a market order, it is executed 'at best', even if it was posted by a LF traders several cycles ago, intending to trade at a different price. This situation shows the potential problem LF traders face in a market populated by their HF counterparts. In cases of a limit order, they risk either to quote a less than competitive order, that is, deep in the book, and not at the top of it as it was their original intention, or to quote an order detached from the rest of the limit orders. If instead they post a market order, it may be executed at an unknown, and potentially worse than intended, price, depending on how the price moves in the meantime. The slow trader's intention was to trade at a price shown when it made the decision to operate - and not at the one available when its order eventually hits the book. For sake of clarification, let us suppose a bid at 100 and an ask at 101. A human trader decides to quote a bid limit order at 100 but during the time it takes its order to reach the exchange server, the market moves to 101 bid-102 ask. The human order is not at the top of the book: not an ideal situation for the slow trader but no loss occurs. In case of 100 bid-101 ask and a buy market ('at best') order at 101, if during the blinking time the market moves up (101 bid-102 ask), then the market order executes at 102, definitely not what was intended by the trader! It is easy to understand how in this situation the latency may exacerbate market volatility, even against the investors' will. If instead the market moves down, 99 bid-100 ask, the market order gets executed at 100, better than intended but still at the prevailing market price. Yet, this situation can even depict a favourable outcome for the slow investor: if the new price was caused by a market swing and the price soon rebounds back to the original level, the slow trader might find itself trading at a profit but this would just be a matter of luck. Probably not many traders would happily face such a double-edged situation. So far it has been described what can potentially happen. Running the simulation shows whether, or not, that actually happens. The pseudo-code with a detailed explanation can be found in Appendix A.

4.3.3. INITIAL DATA

The initial order book is created by adapting the data taken from Labuszewski (2010) at page 52 (Limit Order Book for Sep-09 E-mini S&P 500 Futures, at 9:45AM on August 20, 2009), with the only exception that the liquidity at the fourth price level in the ask book, which was too high with respect to the other values, has been artificially reduced to the average of the liquidity for the remaining price levels. Table 2 shows the original table and the adapted one. In order to simplify the simulation, the number of securities that can fit at each price level has been set equal to a discrete number of lots variable between 1 and the chosen level of maximum liquidity, which was arbitrarily set to 5, but which could have been set to any other number.

Bid Price	**Bid Qty**	**Ask Price**	**Ask Qty**	**Adapted Ask Qty**
1004.00	707	1004.25	588	588
1003.75	1614	1004.50	1838	1838
1003.50	1684	1004.75	2147	2147
1003.25	1899	1005.00	**3637**	**1601**
1003.00	1289	1002.25	1592	1592
1002.75	1029	1005.50	1824	1824
1002.50	1045	1005.75	1386	1386

Bid Price	Bid Qty	Ask Price	Ask Qty	Adapted Ask Qty
1002.25	1980	1006,00	1825	1825
1002.00	1043	1006.25	1712	1712
1001.75	1359	1006.50	1495	1495

Table 2. Adaptation of liquidity in the sample data

A new limit order adds one lot and each trade consumes the same. Overall, this should not affect the realism of the simulation's outcome. The actual liquidity (called Qty in Table 2) at each price level was divided by the liquidity at the top and the result was taken as the mid-value of the lots (3 in case of maximum allowed liquidity of 5 lots). The other lots are proportional to the first one.

4.4. RESULTS AND DISCUSSION

The core of this research consisted in launching the simulation under the two cases, the base case and the trend case, under five different configurations each: scenario 1 - LFT only (QR=0), and scenario 2 - HF-to-LF quoting ratios equal to 1, 2, 20 and 200. A quoting ratio QR=N means that for every quote posted by a LF trader, chances are that N quotes are posted by their HF counterparts. In both the base and the trend cases, the QR=1 and QR=2 yielded a trading ratio falling within a realistic range (50%-70%). The results are shown in Table 3. In both the Base and the Trend case, Table 3 shows the HF-to-LF quoting ratio, where the case QR=0 means LFT quoting only; the Trading ratio shows the actual HF-to-LF trading ratio achieved for the corresponding quoting ratio; MIN and MAX prices represent, respectively, the averages of the minimum and maximum price at which the trades occurred over the 100 cycles of the simulation. It does not matter whether the trend is upward or downward, the difference between the highest and lowest price indicates how volatile the price is. In the literature, volatility is often computed as the standard deviation of a security’s price changes. Yet, in this context, volatility is defined as the average of the differences between the maximum and minimum price reached during each of the 100 cycles of the simulation. In a mathematical formula:

$$\text{Volatility} = \frac{1}{n}\sum_{i=1}^{n}[MAX(cycle\ i) - MIN(cycle\ i)]$$

This is consistent, among others, with Bollen and Whaley (2015), who “compute several range-based estimators of realized volatility, which use only the open, high, low, and closing prices observed in a trading day” (ibid. p.432) and with Hendershott, Jones and Menkveld (2011), who construct daily volatility estimates “based on the daily price range (high minus low)” (ibid. p.19).

Panel A

	BASECASE				
	QR = 0	QR = 1	QR = 2	QR = 20	QR = 200
Trading ratio	0.00%	51.60%	68.03%	95.55%	99.54%
Avg MIN price	1002.7600	1003.5925	1002.3800	1002.6575	1002.6625
Avg MAX price	1005.7225	1005.4925	1005.3800	1005.7675	1005.7600
Volatility	2.96250	2.90000	3.00000	3.11000	3.09750
Std Dev	0.80668	0.81881	0.83182	0.91654	0.87969
Z-score		-0.54375	0.32363	1.20805	1.13107

Panel B

	TRENDCASE				
	QR = 0	QR = 1	QR = 2	QR = 20	QR = 200
Trading ratio	0.00%	50.62%	67.32%	95.03%	98.99%
Avg MIN price	1001.5900	996.3300	998.5800	1001.4075	1001.9875
Avg MAX price	1006.1425	1010.2550	1007.9225	1006.5875	1006.4000
Volatility	4.55250	13.92500	9.34250	5.18000	4.41250
StdDev	1.22953	2.02057	1.79036	1.21962	1.09023
Z-score		39.62572***	22.05452***	3.62335***	-0.85196

Table 3. Results of the simulations

StdDev is the standard deviation for the sample of the Volatility:

$$\text{StdDev} = \sqrt{\frac{\sum_{1}^{n}(x_i - \overline{x})^2}{n-1}}$$

The reason for using the standard deviation of the sample rather than the population is that the outcome of a simulation does not represent the entire population of the trades but only a part thereof.

Z-score represents the value obtained testing the difference between the mean price at the corresponding quoting ratio (i.e., QR=1, QR=2, QR=20, QR=200) and QR=0. Asterisks near the value of Z-score represent the confidence level: no asterisks means that the null hypothesis cannot be rejected at 90% confidence level or below; one asterisk (*) means that the null hypothesis can be rejected at 90% confidence level but not at 95%, two asterisks (**) mean that it can be rejected at 95% but cannot at 99% confidence level, whereas for values marked with three asterisks (***) the null hypothesis can be rejected at 99% confidence level.

In the Base Case (panel A), the null hypothesis of average volatility being equal for QR>0 and QR=0, cannot be rejected for any trading ratio at a confidence level of 90%. The Trend Case (panel B) is different. For QR=1 and QR=2 the null hypothesis is to be rejected at virtually any confidence level (i.e. it is statistically certain that the volatility is higher when HF traders participate). Instead, in case of QR=20 the null hypothesis cannot be rejected at 99.99% confidence level (although it can at 99% level) and for QR=200 it cannot be rejected at as low as 90% confidence level. The quoting ratio can increase for two reasons: the HFT speed increases or (more likely) there are more HF traders taking part. With twice as many HF traders (QR=2 compared to QR=1), the total number of traders does not change appreciably (because of their small number in comparison to the total) but their activity increases. This means that, as the market approaches to one with equal conditions for all (either all fast or all slow), the differential latencies reduce and so does volatility. Examples are provided by the results for QR=20 and QR=200, leading to HF-to-LF trading ratios of about 95% and 99%, respectively, when the volatility is again comparable (and not statistically distinguishable) from the case QR=0. This result seems to strengthen the impact of HFT on volatility even further by providing a counter-proof. Moreover, the findings match intuition. As long as the market is relatively calm, the higher relative speed of some participants with respect to the others does not harm the global functioning of the market itself, and volatility in particular, as even a market order suffering from latency will still find a relatively stable price situation. HFT might perhaps penalise the slower traders but this does not increase market volatility as such. At the contrary, when the market is under pressure, slower traders read a price and, by the time their orders reach the exchange server, the price may already have moved (because in troubled times with HF traders participating, prices move fast). The result is that the price may, involuntarily, be pushed further away as LF traders' market orders were intended to occur at a more favourable (and less volatile) price than the one at which it actually executes. When most trading is carried out by traders experiencing the same low latency (as in the case of QR=20 or QR=200), the market tends to show less volatility because of the queueing effect, which potentially penalises LF traders' market orders is less pronounced as they trade much less. In a homogeneously fast trader environment (when QR is high) the volatility does not increase with respect to the scenario with LF traders only.

4.5. CONCLUSION

This chapter presented two simulation cases (the Base Case and the Trend Case) in order to investigate the impact of HFT on volatility (one scenario assumes HFT activity whereas the other does not). If the market is quiet (this is simulated in the Base Case), then no extra volatility appears when High-Frequency traders enter the game with respect to a scenario with slow participants only. However, when the market is under pressure (as in the Trend Case), the simulation found a statistically significant and very strong impact of High-Frequency Trading on volatility. The latency experienced by slow players causes their market orders to execute with a relatively high delay, compared to the instantaneous trading executed by HF traders. This delay leads to price execution uncertainty and, under stressed market conditions, the price volatility may be further exacerbated by this delay. As the number of HF traders increases, the market tends to become homogeneous again, with most participants experiencing comparable latencies, and therefore volatility returns to usual values. The results of these simulations confirm the findings of Kirilenko et al. (2011) on the Flash Crash, who in their conclusion section state that "the trading of HFTs, appears to have exacerbated the downward move in prices" (ibid. p.3). Chapter 7 will inspect exchange data to verify the presence of the phenomenon found via the simulation.

5. DOES HIGH-FREQUENCY TRADING CREATE A TWO-TIER MARKET?

5.1. INTRODUCTION

The adoption of High-Frequency Trading practices in the financial markets has brought great changes and some concern, as found by Beunza, Millo and Pardo Guerra (2012), who carried out several interviews with senior market participants. Many aspects of HFT have been deeply investigated but no conclusive understanding of its impact on the markets has been reached so far. However, some authors have highlighted possible risks implicit in HFT strategies and practices, although not always detailing or quantifying them. A typical example is what can be called market tiering. By this it is meant the tendency of the financial markets to split into two communities, to some extent only dealing internally to the community itself. Alternatively, the same concept can be viewed as a two-lane market: a fast lane, where trading enjoys tight bid-ask spread, and likely profitable deals, and a slow lane, where participants must content themselves with the leftover, hoping for middle- or long-term profits, that are inherently riskier. The mechanism based on which market tiering can appear is strongly linked to the wider topic of non-linearity: at a certain stage the behaviour of a dependent variable that had, until then, changed linearly with the change in the independent variable, abruptly breaks that kind of relation and starts reacting unpredictably. The purpose of the simulation presented in this chapter is contributing to a quantitative measurement of that tendency, if it exists at all. In order to do so, the simulation develops an Agent-Base Model, implemented via a software routine that mimics the behaviour of several market participants. They quote limit orders and post market orders under certain assumptions that will be discussed thoroughly in the following sections. The aim of the simulation is to replicate, although with some simplification, how a real market works. All trading is recorded onto a database and the outcome analysed to verify whether or not the tiering tendency actually occurs.

The next paragraph presents the concepts of non-linearity and phase change as it could be applied to market tiering. Then it follows the description of the methodology (paragraph 5.3), the explanation of the simulation (paragraph 5.4) and the results are presented (paragraph 5.5) and discussed (paragraph 5.6). Paragraph '5.7. Conclusion' terminates the chapter.

5.2. IS MARKET TIERING A THREAT TO FINANCIAL STABILITY?

5.2.1. ON SPEED, AGAIN

One of the main features (among others) characterising HF traders is their relative speed compared to other, low-frequency, traders. In today's market environment, speed is so critical that any limit to it is by many considered harmful to the correct functioning of the whole financial system. According to the special relativity theory, the speed of light provides an upper limit to the speed any physical particle can reach; even information transfer is bound to that limit. In practice, the most advanced electronic communications technology is able to reach only about 2/3 the speed of light because of networking, routing and other practical limitations. In order to understand the impact speed has on the financial markets let us consider a trader in Frankfurt who wishes to sell a security via a market order to the NYSE in New York. The trader needs to get the information about the bid price of that security, then it has to work out whether, in its view, the limit order price is favourable and, if so, it will send its market order across the Atlantic. For sake of simplicity let us assume the time spent for making the trade decision to be zero, resulting in a latency for the whole process equal to the price transmission from New York to Frankfurt plus the order transmission from Frankfurt back to NY. The distance between the two cities is about six thousand kilometres, which means 12,000 km for the return trip. The speed at which information may travel in practice is around 200,000 km/sec. This results in a roundtrip time of 0.06 seconds. To understand how much six hundredths of a seconds is, suffice it to say that the price of the E-mini S&P 500 Futures contract on the day of the Flash Crash around 2:45:28pm moved 25 ticks downward in just 7 milliseconds (from 1062.25 index points at 2:45:28.107, down to 1056 at 2:45:28.114) for a total of 6.25 index point drop. That means a $312.50 delta per each E-mini S&P 500 futures contract. Admittedly, the 2pm to 3pm of May 6, 2010 was a very frenetic time and it cannot be taken in any way as representative of the typical trading day. At another time on the same day, in 0.06 seconds the price just moved one tick downward while in the meantime the book was updated 155 times. But even during a quiet period, in six hundredths of a seconds the price of a security may well change a few ticks in one direction or the other, with the result that the price an investor intended to buy or sell may no longer apply when, an eye-blink time later, its order gets executed. Angel (2014) borrows from the special relativity theory to notice that every venue can only have certainty about its local prices, whereas, just because information cannot travel faster than the light, prices of some venues might not be realistic. "At the same moment in time, two geographically separated participants may observe two different 'best' prices" (ibid. p.5). The same concept is also expressed by Haldane (2011), who points out that before a low-frequency trader is able to execute its own order, its high-frequency, co-located, counterparts may have executed many thousands - and therefore the price may have changed. This is definitely a source of concern for several reasons linked to the orderly functioning of the markets and to financial stability: fast traders might front-run the slower ones [as illustrated, among others, by Lewis (2014a)]; the practice of rapid order quoting and cancelling creates the so-called ghost liquidity [Zhang and Baden Powell (2011)]; it may change prices between order generation and order execution (as in the Frankfurt to NY example above) leading LF traders to buy higher or to sell lower than otherwise they would. A human blink lasts about 150ms; events below that threshold are negligible to humans. Johnson et al. (2013) notice how one of the quickest, albeit complex, activities human beings have proved to handle well is when a chess Grandmaster realises her own king is in checkmate. This takes no less than 650ms, sometimes more (according to the complexity of the position on the chessboard). A human-made buy or sell decision, even in the most unambiguous scenario, is unlikely to take less time, and will probably take more. The difference between HFT and human trading is made even more pronounced by the communication latency advantages widely used in today's trading direct data feeds, dedicated networks, and co-location.

5.2.2. TWO MARKETS?

This situation has not passed unnoticed. The US Sen. Edward Kaufmann, as reported by Advent (2012), voiced concern about a two-tiered market: "one market for huge-volume, high-speed players, who can take advantage of every loophole for profit, and another market for retail investors, whose orders are seemingly filled as an afterthought" (ibid. p.9). In an interview reported in Friederich and Payne (2012a), a director of MTF Turquoise went even further by stating "Where a customer is, relative to the different market centres, means that their view of price [...] is totally different to a customer located somewhere else. I think we have to abandon this idea that there is a universal truth for the best currently available price" (ibid. p.25). This means that the better price shown by exchange A may not yet be known by everyone at the time an order is executed at exchange B. As it is often the case in human matters, people are split among the haves and the have-nots. Lewis (2014a) puts it clearly: "The haves paid for nanoseconds, the have-nots had no idea that a nanosecond had value. The haves enjoyed a perfect view of the market, the have-nots never saw the market at all" (ibid. p.39). One could argue that in the market there have always been faster, or better-located, traders than others. Even in an open outcry environment the nearer sitting and faster traders got the advantage. MacKenzie (2015) notices how a trading pit was shaped like an amphitheatre with many traders and brokers crowded together, all able to see and hear one another. The crucial difference with HFT environment is that, at that time, the less well-placed traders were at least able to observe what was going on whereas today they may not even have a clue of what is happening behind the scenes because things occur far too fast to be noticed by them. "High-speed traders in earlier centuries employed communication mediums that were faster than the norm at the time. Such devices have included fast sailing ships, courier pigeons, express coaches, smoke and hand signals, semaphore flags, mirrors, the telegraph and private telephone lines" [Markham (2015), p.3]. Whereas it is still true today that the faster trader wins, the slower traders not only lose, but they will not even receive the same information allowing them to take part to the competition. The playground is definitely uphill for them. Trading at best prices occurs among the ignorance of most traders. This phenomenon, which is likely to create a two-tier market, is further exacerbated by the common practice of exchanges selling data feeds in advance to the officially distributed information. Some exchanges sell market data directly to their subscribers, when other investors learn about market events from the tape a few milliseconds later - and others from the Wall Street Journal on the following day. This is a consequence of Rule 603(a) of RegNMS (2005). The 'fair and reasonable' and 'not unreasonably discriminatory' requirements set by the Rule about distribution of market information have been widely accepted as prohibiting exchanges to transmit data to a subscriber any sooner than the same data is transmitted to the general public via the Securities Information Processor (SIP). However, because of the latency inherent in the process of sending data via the SIP, sufficiently fast traders may find it more convenient to build their own National Best Bid Offer (NBBO) rather than using the one provided by the official market data consolidator. This possibility, widely applied in the real world, also seems to confirm the existence of a two-tier market, where in tier 1 sit all the aware investors who receive information in real time - and can act upon it. Given the speed of the tier 1 investors, whether other investors come to know price information a few milliseconds later, the following day, or ignore it altogether is, under any practical respect, largely immaterial. The New York Times, on July 23, 2009 announced that "rather than being shown to all potential sellers at the same time, some [...] orders were most likely routed to a collection of high-frequency traders for just 30 milliseconds — 0.03 seconds — in what are known as flash orders. While markets are supposed to ensure transparency by showing orders to everyone simultaneously, a loophole in regulations allows marketplaces like Nasdaq to show traders some orders ahead of everyone else in exchange for a fee". One may argue that 30ms is such a short time that should affect no one, but the same person would be (rightly) outraged if the delay would be 30 seconds or 30 minutes instead. Yet, conceptually there is no difference. HF traders have all the time to exploit short timeframe informational advantage to the detriment of slower traders. This point is also raised in Pandey and Wu (2015), according to whom "some high frequency traders

(HFTs) are able to glean order flow information ahead of other traders and profit from it. The current arrangement is beneficial for HFTs and exchanges, but detrimental for investors" (ibid. p.53). A human is simply not fast enough to react timely to market events which may instead be considered appealing by a HF trader. The space in which the former operates is the intuition of future profitability whereas the space of the latter is the very short-term market movements - two completely different dimensions. It must be noticed, however, that failure to compete on speed does not prevent traditional traders from profitably operating on the market: if their medium- or long-term financial analysis proves correct, they will make money nonetheless. Yet, the risk level implicit in their trading is clearly different from the one faced by a HF trader who just needs to wait for a very tiny price oscillation in the favourable direction to profitably closing its position.

5.2.3. NON-LINEARITIES

The practice of selling price data worsens market quality according to Easley, Lopez de Prado and O'Hara (2016): it increases volatility, decreases liquidity, discourages the production of fundamental information and makes the markets less efficient compared to situations when all traders observe the same prices at the same time. In other words, it negatively affects financial stability. Years before, Easley, Lopez de Prado and O'Hara (2012) had already highlighted this difference: "Financial analysts' conferences are one milieu where low-frequency traders (LFT) converse on subjects as broad and complex as monetary policy, asset allocation, stock valuations, financial statement analysis, and the like. HFT conferences are reunions where computer scientists meet to discuss TCP/IP connections, machine learning, numeric algorithms to determine the position of an order in a queue, the newest low-latency co-location architecture, game theory, and most important of all, the latest variations to exchanges' matching engines. One would conclude, correctly, that the LFTs and the HFTs seem worlds apart" (ibid. p. 20). The Foresight Report (2012) summarises the matter: "It seems unlikely that the future of CBT [Computer-Based Trading] in the financial markets leads merely to a faster system, and therefore to more frequent crashes and crises, purely on the (metaphorical) basis that 'the same old movie' is now being played at fast-forward speed. Rather, it seems more likely that, despite all the benefits, CBT has the potential to lead to a qualitatively different and more obviously nonlinear financial system in which crises and critical events are more likely to occur in the first place" (ibid. p.73). On the same wavelength is O'Hara (2015): "From the way traders trade, to the way markets are structured, to the way liquidity and price discovery arise – all are now different in the high frequency world" (ibid. p.257). These statements raise a fundamental point. If the elements of a system increase their speed, under a linear regime one could expect that the system's behaviour remains qualitatively the same, just faster. But the outcome of a non-linear system, undergoing the same acceleration is not as predictable. Particularly worrying is the Millennium Bridge case, both from the engineering and financial viewpoint. Since many years, engineers have learnt how to take into account the resonance effects into bridge design and in particular they warned about harmonised frequencies, like the march of soldiers, potentially leading to collapse. Indeed, when soldier cross a bridge they are asked to break step in order to avoid applying harmonised frequencies to the structure. As reported by Danielsson (2013) "the Millennium Bridge was supposed to sway gently in response to the Thames breeze. A gust of wind - an exogenous shock - hit the bridge, causing it to move sideways and wobble. When this happens a natural reaction is to adjust one's stance to regain balance. By doing so, the bridge gets pushed back, making it sway even more, causing people to adjust their stance yet again - more and more at the same time - this time pushing the bridge in the opposite direction" (ibid. p.42). The bridge was closed a few days later and only reopened in 2002. The triggering factor was a relatively strong wind but it would not cause any problem if the people were not moving synchronously to regain their balance. The wobble continued long after the triggering gust of wind had terminated. That was a typical case of self-reinforcing feedback loop. "The ultimate lesson from the Millennium Bridge is that it is not the shock that matters but the feedback mechanism that allows a small shock to be amplified into a large event" (ibid. p.42). It is worth noticing that neither event (the gust of wind or the pedestrians crossing the bridge) in isolation would be capable to cause the event: it is their combination that highlighted a weakness in the system design. The combination of individually innocuous effects, as explained by the Theory of Chaos, is a major risk factor for complex systems. Financial markets, as complex systems, may actually behave like other complex systems. They might obey to the same laws of non-linear behaviour, and once they are subject to perturbation beyond the linear-to-non-linear phase transition, they may start behaving unpredictably. Non-linear phenomena tend to arise after a certain threshold has been trespassed: in such cases we talk about phase transition. As recognised by Sornette (2003) and Danielsson (2013), phase transitions may also happen in financial markets. Phase transition is not easy to define in terms of financial markets as they do not display easy-to-grasp features like, for example, solid state for ice, liquid for water and gaseous for steam. However, the physical appearance of H_2O that we are used to is an incidental consequence of its molecules' microscopic structure; it is actually at that microscopic level that phase transitions occur. Similarly, phase transitions in financial markets may well affect aspects difficult to grasp without appropriate theoretical or practical tools. Exactly like many mechanical systems, under certain conditions feedback loops may arise - and at very high speed such loops may display a non-linear relationship with respect to their behaviour at lower speed and/or with initial conditions. Doubling the speed of the system does not necessarily lead to just doubling the speed of its effects. Water slowly flowing in a pipe shows a laminar movement but increasing its speed beyond a certain level causes the movement to turn into a turbulent one, displaying chaotic features and vortices (non-laminar formations) appearing randomly. On the one side an environment more sensitive to small perturbations, amplified by speed, is more prone to systemic instability; the 'butterfly effect' is a factor that can no longer be underestimated. On the other side, the somehow esoteric view of the 'invisible hand' which has the collateral effect of raising the wealth of the whole community while the individual seeks one's own benefit may now come under questioning.

5.3. METHODOLOGY

Based on the academic findings described above, it looks like there is a possibility that entrance of HF traders to the market causes some sort of phase transition, the outcome being a split of the market into two tiers with little interaction to one another: the fast lane, where high-speed computers and networks run against each other - and the slow lane suitable for latency-prone computers and humans. The rationale behind this scenario is that HF traders tend to post market orders when the spread is thin (in the simulation the threshold has been set to one tick, where the tick is the minimum price change allowed by the exchange rules) and mostly limit orders otherwise. Hagströmer, Nordén and Zhang (2014) confirm this approach stating that HF traders submit often market orders when liquidity is abundant (and therefore spread is thin) and more limit orders when the bid-ask spread is wide. Hendershott and Riordan (2013) find that "[w]hen spreads are narrow ATs [algorithmic traders] are less likely to submit new orders, less likely to cancel their orders, and more likely to initiate trades" (ibid. p.1001). This intuitively leads to a tight market populated by aggressive orders coming from HF traders and a market featuring wider spread in which HF traders post limit orders, leaving mostly LF traders to play aggressively. Despite the topic of market tiering and phase transition have been approached by some authors, it looks like not many quantitative studies have been carried out to conclude that HFT actually causes the market to create two separate airlocks, one for fast and one for slow traders. Whereas some studies support the view that HF traders tend to mostly deal with counterparts operating at similar speed, not much effort has been dedicated to identify a similar tendency of LF traders to mainly have other slow traders as counterparts. In order to understand better whether or not that intuition is compatible with a concrete scheme, this research simulated several scenarios, with the goal to replicate a realistic behaviour of both HF and LF traders. The scenarios are different in only two aspects, namely, speed and how much HF traders act with respect to the other market participants. Since the main discriminant between the two categories of traders is speed, non-HF traders are collectively viewed as low-frequency traders, irrespective of their speed relative to each other, their individual strategies, whether market makers or takers, or any other specific features characterizing them. All the differences are implicitly there, albeit hidden in the random choice of the three dimensions: the book to act upon (bid or ask), the kind of participant (LF or HF), and the order type (limit or market). A parameter is set at the beginning of each simulation launch: it represents the quoting ratio (QR) between HF traders and low-frequency participants. The higher the value of QR, the higher the probability of a HF trader being selected by the random-choosing algorithm. The selected participant shall either more likely quote a price or post a market order according to the width of the spread. The testable hypothesis of this simulation is whether the participation of HF traders increases the likelihood of both HF and LF traders to mostly deal, respectively, with fast and slow counterparts.

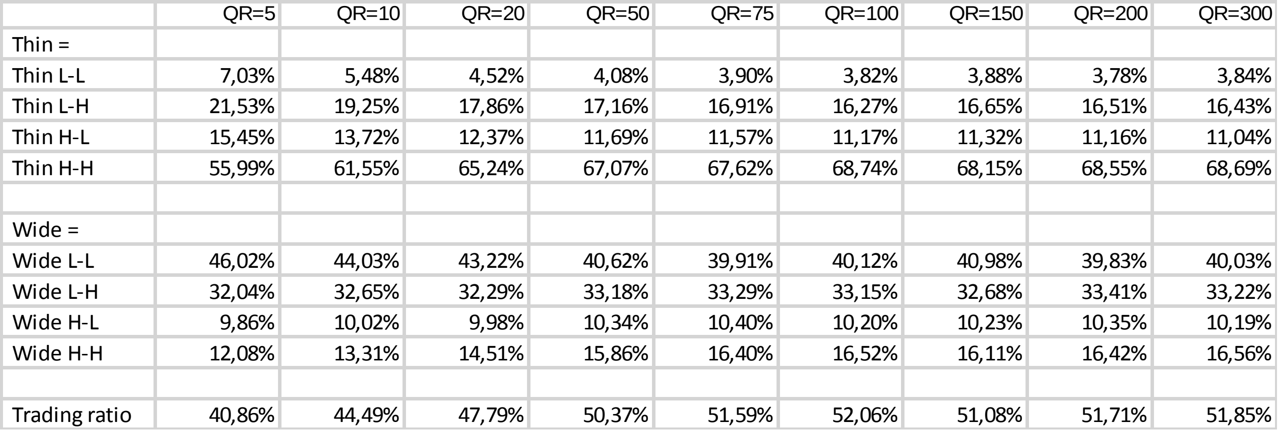
5.4. DESCRIPTION OF THE SIMULATION

The main routine is repeated 100 times in order to provide sufficient data for statistical purposes. At the end of each repetition, the resulting trading data is saved onto the database, split between trading at thin spread (one tick) or wide spread (more than one tick) and category of aggressive and passive parties (L for low-frequency trader, H for high frequency trader), in order to apply statistical analysis. The main routine runs over 10,000 cycles, each cycle representing a time period, whatever its length, during which either a limit or a market order is executed. The first operation the main routine performs is randomly selecting the book it will act upon: either bid or ask. Then, if the bid-ask spread is larger than one tick, HF traders have less chances to be (randomly) selected than LF traders. The rationale for this choice is the preference of HF participants for tight markets. Conversely, if the spread is equal to one tick, then whether an HF or a LF traders operates, is (randomly) decided by an algorithm that takes the QR parameter into account. For example, with QR=1 both type of traders have the same chances to be selected whereas with QR=10, a HF trader is ten times more likely to be selected than a low-frequency counterpart. Since HF traders regard wide spread too risky for their ultra-short-term strategy, when the spread is wide higher probability is given to selection of limit order than to market orders. A simplifying assumption about liquidity has been made. All orders are supposed to only handle one lot of securities. A limit order will quote one lot at the given price and a market order will only trade one lot. The maximum liquidity allowed at any price level is 5 (although this is a parameter that can be changed). The order type is also selected using a random-number generation algorithm, yet assigning more chances to market order if the liquidity is in the upper half of the allowed range (i.e. if 3 through 5 limit orders are present at that price level). The order is then applied to the book selected earlier. However, in order to simulate quotes at a better price than top of book a special feature has been implemented. As described above, every limit order adds one lot to the book and every market order consumes exactly one lot; moreover each price level can only reach a maximum liquidity (set at initialisation time) and never trespassing it. So, when the top-of-book price level has reached the maximum liquidity, this is taken as sufficient consensus by the market on the soundness of that price to suggest the investor quoting the next price level. The simulation displays a moderately elastic behaviour: when the spread is tight and liquidity is high, preference is given to market orders, so widening the spread; instead, when the spread is large, the random number generation algorithm gives preference to limit orders, closing the gap again. This behaviour complies with the findings of Carrion (2013): "HFTs provide liquidity when it is scarce and consume liquidity when it is plentiful" (ibid. p.710) and Jarnecic and Snape (2014), who find HF traders resolving temporal liquidity imbalances. The pseudo-code with a detailed explanation can be found in Appendix B.

5.5. RESULTS OF THE SIMULATION

5.5.1. MAIN APPROACH

The simulation has initially been run in three different scenarios, depending on the quoting ratio between HF orders and other traders' orders. The target value was compatible with the trading ratio found in the literature [CFCT-SEC (2010a), Cliff, Brown and Treleaven (2010), Friederich and Payne (2011), Haldane (2011), Ahlstedt and Villyson (2012), Foucault (2013)]. Since the simulation does not allow to set the trading ratio directly, but only the quoting ratio, a few attempts needed to be made, changing the QR every time, until an acceptable value for the trading ratio was eventually achieved. Two cases were investigated: bid-ask spread = 1 tick (thin spread) and bid-ask spread > 1 tick (wide spread). The simulation has been run several times with QR values ranging from 1 to 300. All quoting ratios ranging from 5 through 300 yielded an acceptable trading ratio (40-60%), as shown in table 4.



	QR=5	QR=10	QR=20	QR=50	QR=75	QR=100	QR=150	QR=200	QR=300
Thin =									
Thin L-L	7,03%	5,48%	4,52%	4,08%	3,90%	3,82%	3,88%	3,78%	3,84%
Thin L-H	21,53%	19,25%	17,86%	17,16%	16,91%	16,27%	16,65%	16,51%	16,43%
Thin H-L	15,45%	13,72%	12,37%	11,69%	11,57%	11,17%	11,32%	11,16%	11,04%
Thin H-H	55,99%	61,55%	65,24%	67,07%	67,62%	68,74%	68,15%	68,55%	68,69%
Wide =									
Wide L-L	46,02%	44,03%	43,22%	40,62%	39,91%	40,12%	40,98%	39,83%	40,03%
Wide L-H	32,04%	32,65%	32,29%	33,18%	33,29%	33,15%	32,68%	33,41%	33,22%
Wide H-L	9,86%	10,02%	9,98%	10,34%	10,40%	10,20%	10,23%	10,35%	10,19%
Wide H-H	12,08%	13,31%	14,51%	15,86%	16,40%	16,52%	16,11%	16,42%	16,56%
Trading ratio	40,86%	44,49%	47,79%	50,37%	51,59%	52,06%	51,08%	51,71%	51,85%

Table 4. Trading between HF and LF traders for thin spread and for wide spread

The rows under 'Thin=' display the percentage of trades occurred with a bid-ask spread equal to one tick while under the row 'Wide=' are shown the percentage of trades with a spread greater than one tick. Underneath are reported the percentage of trades occurred between two Low Frequency traders (L-L), between an aggressive LF trader and a passive HF trader (L-H), the other way round (H-L), and the trading between two HF traders (H-H). The row 'Trading ratio' shows the actual ratio between HFT and LFT, depending on the QR. The trading ratio initially grows and then stabilises, probably in correspondence of some sort of 'saturation' of the HFT activity. As table 4 shows, in case of thin spread, as soon as the trading ratio trespasses the 40% threshold (column QR=5), the percentage of trading occurring among HF traders (row 'Thin H-H') goes from 55% of all trading with thin spread, to nearly 69% (column QR=300). Figure 7 (QR on the x-axis) provides a graphical representation of the data contained in table 4. The percentages on the x-axis indicate the trading ratio obtained, corresponding to the QR ranging from 5 through 300. The dark line shows the trading occurred between two HF traders with spread equal to one tick, whereas the light line shows the trading between LF traders with spread greater than one. The results of the simulation seem to confirm that the more HF traders act on the market with thin spread, and the more they tend to trade among themselves. This depicts a rather clear-cut situation: the higher the trading ratio and the more HF traders deal among themselves. As a counter-proof, it can be noticed that the percentage of mixed trading (rows 'Thin L-H' and 'Thin H-L' added together) declines from 37% (column QR=5) to less than 28% (column QR=300). The case of wide spread is less clear. The results on the left-hand side (columns with QR up to 20 in table 4) show a relative predominance of trading among Low-Frequency participants, while for QR=30 and above mixed trading (Wide L-H plus Wide H-L) surpasses the Wide L-L trading. Table 5 computes the statistical significance of the equality between the two means (with independent samples). It is interesting to notice that the amount of LF-to-HF and HF-to-LF trading in case of wide spread is quite high in all scenarios, ranging between nearly 42% and nearly 44%. This is consistent with the findings of the UK's Financial Conduct Authority [Aquilina and Ysusi (2016)], which reports 43% of all trades occurring across categories.

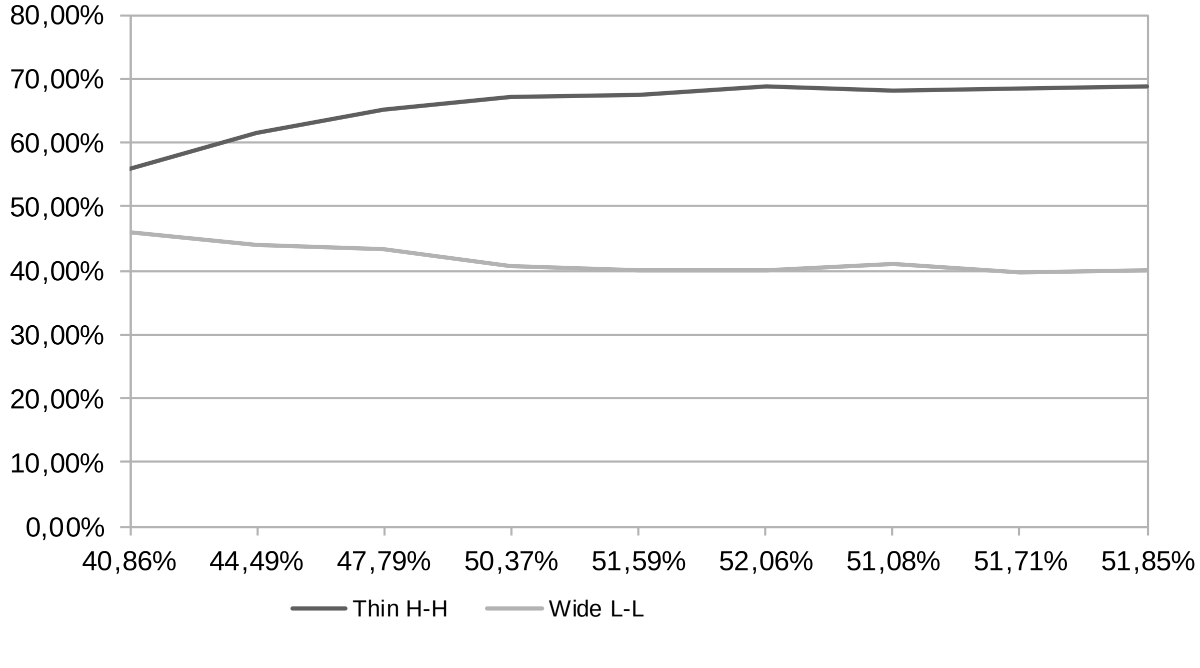
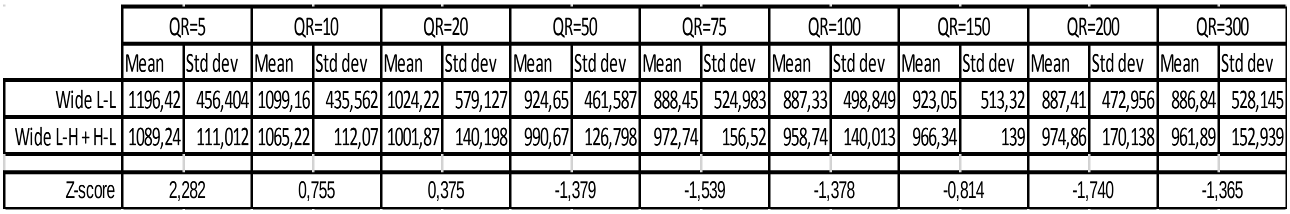


Figure 7. Percentage of trading among High-Frequency traders for thin spread and among Low-Frequency traders for wide spread

As shown in table 5, the Z-score for each QR allows to easily state that the L-L trading is significantly different, at 95% level of confidence, from the sum of L-H and H-L trading together only for QR=5 (the null hypothesis can therefore be rejected for that QR). In all other scenarios the null hypothesis of equality between L-L and mixed trading at 95% level cannot be rejected. The intra-community effect is less evident for LF traders when the spread is wide: the communities' tendency to trade with each other seems more stable.



	QR=5		QR=10		QR=20		QR=50		QR=75		QR=100		QR=150		QR=200		QR=300	
	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
Wide L-L	1196,42	456,404	1099,16	435,562	1024,22	579,127	924,65	461,587	888,45	524,983	887,33	498,849	923,05	513,32	887,41	472,956	886,84	528,145
Wide L-H + H-L	1089,24	111,012	1065,22	112,07	1001,87	140,198	990,67	126,798	972,74	156,52	958,74	140,013	966,34	139	974,86	170,138	961,89	152,939
Z-score	2,282		0,755		0,375		-1,379		-1,539		-1,378		-0,814		-1,740		-1,365	

Table 5. Difference between Wide L-L and Wide mixed trading

The results obtained for both thin spread and wide spread do not show a clear phase transition point; it looks like a smooth convergence toward a situation in which between 57% and 69% of trading with tight spread occurs among HF traders and a more stable percentage of trading with wide spread goes among LF traders (between 40 and 46 percent). This means that there is a clear sign of a tiering market with thin spread and only a much weaker evidence in the case of wide spread. The percentage of Thin H-H trading is systematically higher than 50%, whereas the Wide L-L trading percentage, despite being higher than the individual percentages of L-H, H-L or H-H trading, never reaches the 50% threshold. The graphic in figure 7 confirms the 'progressive response rather than a step change', as found by Cartlidge and Cliff (2012) for High-Frequency traders dealing more among each other as their trading activity increases.

5.5.2. ALTERNATIVE APPROACH

Financial markets are complex systems and in order to try and isolate one specific cause-effect relationship within a simulation, it is important to minimise the number of possible causes that may have had an impact on the effect under study. In the main approach described above, the main feature implemented was the higher chance of HF traders to act when the spread is tight and their reluctance to cross the spread when it is wide. However, although this sounds quite reasonable, every simplification goes to the detriment of the realism of the simulated system. In particular, there is another major feature that differentiates the behaviour of HF traders to their slower counterparts and vice versa: by definition Low-Frequency traders are slower that HF traders. In the previous simulation this speed difference was implemented as a circular queue in which LF traders' orders are placed for later execution. In other words, an order (either limit or market) coming from a HF trader is executed immediately whereas an order from a LF trader will only be executed after a certain number of cycles (the choice was 650, assuming one cycle on average equal to one millisecond) to simulate the latency due to network-related delays and possibly other factors. The effect is that LF orders relate to market prices that were applicable several cycles before actual execution. Thus, a new set of simulation runs has been launched without this feature implemented, that is, with no latency suffered by LF traders, the difference between the two communities only being the reluctance of the HF traders to post aggressive orders when the spread is wide (as was already the case in the previous approach). Were the result of this approach similar to the previous one, it could mean that the real factor leading to market tiering was the behaviour of the traders rather than their relative speed. The results are reported in table 6. Although in case of thin spread there is a certain amount of trading within the HFT community, that percentage never reaches 50%, even for high trading ratio between HF and LF traders, while it was consistently higher with the previous approach, when the LF traders were actually slower. On the other side, when the spread is wide, the main difference to the previous approach is the more balanced percentages of aggressive trading between the two communities. While in table 4 the LF-initiated trading was always above 70%, in table 6 it ranges between 50% and 60%. Figure 8 shows a graphical representation of the numbers displayed in table 6. The difference between the two approaches is clear when the graphics in figure 7 and in figure 8 are compared. When no speed difference between HF and LF traders is implemented into the simulation, the situation changes significantly, as shown in table 7 and table 8. Again, the difference between two means (with independent samples) was applied. Similar to the main approach described above, the mean and the standard deviation for each level of QR have been displayed for both the thin spread and the wide spread cases. In the thin spread case the HF-to-HF trading was compared against the percentage of LFT versus passive HF liquidity suppliers plus the percentage of aggressive High-Frequency Trading versus passive LF traders, or Thin L-H plus Thin H-L (table 7).

Panel A

	QR=3	QR=4	QR=5	QR=6	QR=7	QR=8	QR=9	QR=10	QR=20	QR=30
Thin =										
Thin L-L	12.93%	11.29%	10.25%	9.70%	9.15%	8.74%	8.57%	8.30%	7.66%	7.31%
Thin L-H	13.80%	13.46%	13.02%	12.77%	12.47%	12.29%	12.49%	12.22%	11.59%	11.31%
Thin H-L	38.82%	37.92%	37.49%	36.88%	36.71%	36.59%	36.27%	35.70%	35.09%	34.69%
Thin H-H	34.45%	37.33%	39.24%	40.65%	41.67%	42.37%	42.67%	43.78%	45.67%	46.69%
Wide =										
Wide L-L	48.97%	48.14%	47.39%	46.81%	46.43%	45.92%	45.72%	45.55%	44.97%	44.48%
Wide L-H	13.87%	13.58%	13.50%	13.58%	13.46%	13.42%	13.71%	13.52%	13.50%	13.46%
Wide H-L	28.93%	29.74%	30.34%	30.57%	30.98%	31.40%	31.27%	31.71%	31.90%	32.27%
Wide H-H	8.23%	8.54%	8.78%	9.04%	9.13%	9.26%	9.30%	9.22%	9.63%	9.78%
Trading ratio	51.17%	52.92%	54.36%	54.99%	56.03%	56.53%	56.39%	57.11%	57.99%	58.84%

Panel B

	QR=40	QR=50	QR=60	QR=70	QR=80	QR=90	QR=100	QR=150	QR=200	QR=250	QR=300
Thin =											
Thin L-L	7.18%	7.00%	7.03%	7.04%	6.98%	6.83%	6.97%	6.89%	6.80%	6.89%	6.76%
Thin L-H	11.39%	11.24%	11.28%	11.37%	11.20%	11.29%	11.27%	11.11%	11.35%	11.14%	11.01%
Thin H-L	34.25%	33.91%	34.13%	34.07%	33.85%	33.74%	33.52%	33.68%	33.74%	33.73%	33.74%
Thin H-H	47.17%	47.85%	47.55%	47.53%	47.97%	48.14%	48.25%	48.32%	48.11%	48.24%	48.49%
Wide =											
Wide L-L	44.86%	44.46%	44.21%	44.82%	43.92%	44.22%	44.34%	44.03%	44.30%	43.96%	43.78%
Wide L-H	13.36%	13.23%	13.35%	13.36%	13.26%	13.09%	13.12%	13.37%	13.32%	13.42%	13.11%
Wide H-L	31.99%	32.48%	32.44%	32.06%	32.65%	32.69%	32.68%	32.67%	32.47%	32.58%	33.01%
Wide H-H	9.80%	9.82%	10.00%	9.75%	10.17%	10.01%	9.87%	9.94%	9.91%	10.04%	10.10%
Trading ratio	58.49%	59.08%	59.20%	58.45%	59.67%	59.58%	59.23%	59.55%	59.16%	59.46%	59.98%

Table 6. Trading between HF and LF traders for thin spread and for wide spread

In the wide spread case (table 8) the comparison is between trading within the Low-Frequency traders community versus mixed trading (Wide L-H and Wide H-L added together). The row 'Z-score' displays the results of the test for the difference between the two means. In the thin spread case, a predominance of the HF-to-HF against mixed trading cannot be noticed for values of QR less than 40. For those values of QR, the Z-score ranges between -25.37 and +1,02, that is to say, from strong dominance of mixed trading to near equality (at 95% level of confidence). For values of QR>=40 the predominance of HF-to-HF trading is statistically significant at 95% level. At the contrary, in case of wide spread, the LF-to-LF trading is only statistically predominant, at 95% level, for low values of QR (when HF traders are less active, Low-Frequency traders take the lion's share), but for QR=9 or higher, the mixed trading becomes indistinguishable, at 95% significance level, and then progressively more and more relevant as QR increases. For QR>=30, mixed trading is significantly predominant compared to LF-to-LF trading.

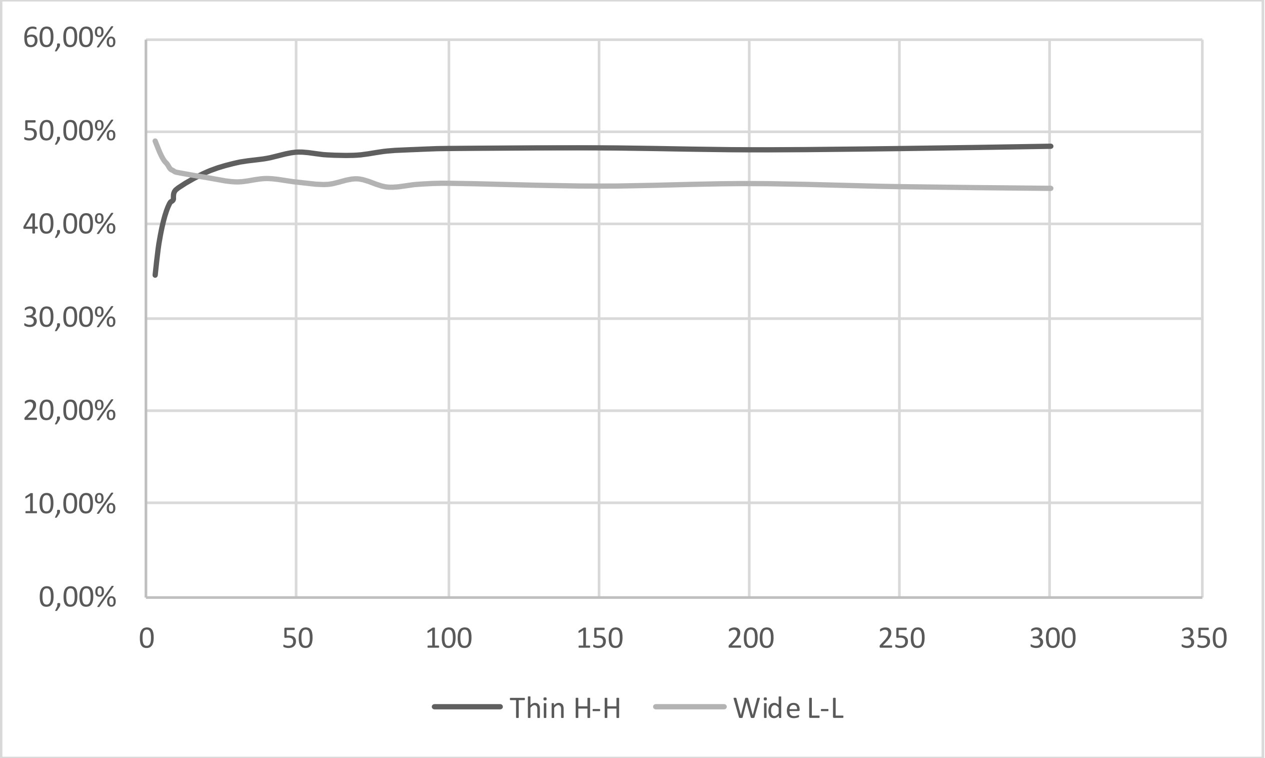


Figure 8. Percentage of trading among High-Frequency traders for thin spread and among Low-Frequency traders for wide spread

Panel A

	QR=3		QR=4		QR=5		QR=6		QR=7		QR=8		QR=9		QR=10		QR=20		QR=30	
	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
Thin H-H	397,47	48,78	438,62	51,65	471,46	55,76	487,89	52,10	513,36	60,54	520,76	59,06	519,94	58,50	544,82	66,60	564,25	69,71	589,20	67,72
Thin mixed	607,22	66,77	603,74	61,35	606,91	58,46	595,94	52,12	605,88	58,14	600,82	50,86	594,15	54,36	596,25	62,32	576,73	56,12	580,48	52,15
Z-score	-25,367		-20,590		-16,767		-14,662		-11,023		-10,271		-9,293		-5,639		-1,394		1,020	

Panel B

	QR=40		QR=50		QR=60		QR=70		QR=80		QR=90		QR=100		QR=150		QR=200		QR=250		QR=300	
	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
Thin H-H	585,56	72,27	599,15	71,34	599,06	68,45	582,84	71,33	612,02	72,29	612,49	68,92	606,98	71,61	612,19	77,81	602,90	78,89	608,85	73,18	617,41	74,14
Thin mixed	566,59	50,99	565,30	55,53	572,10	46,62	557,21	51,78	574,71	52,33	572,85	46,75	563,41	45,52	567,52	51,32	565,10	48,73	566,30	52,19	569,76	50,16
Z-score	2,145		3,744		3,256		2,908		4,181		4,760		5,135		4,792		4,077		4,734		5,323	

Table 7. Difference between the two means (thin spread)

Panel A

	QR=3		QR=4		QR=5		QR=6		QR=7		QR=8		QR=9		QR=10		QR=20		QR=30	
	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
Wide L-L	891,62	72,49	862,38	73,02	835,76	70,95	823,52	62,21	803,08	74,05	797,78	70,81	793,76	68,49	783,78	82,23	768,93	78,94	753,72	77,05
Wide L-H + H-L	779,25	31,85	775,93	33,47	773,15	32,20	776,67	30,46	768,59	32,85	778,66	30,01	780,98	31,42	778,35	30,44	776,33	33,80	774,89	34,94
Z-score	14,192		10,762		8,036		6,764		4,257		2,486		1,696		0,619		-0,862		-2,502	

Panel B

	QR=40		QR=50		QR=60		QR=70		QR=80		QR=90		QR=100		QR=150		QR=200		QR=250		QR=300	
	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
Wide L-L	764,74	77,76	752,59	74,11	747,39	74,99	764,95	75,87	736,54	80,94	743,15	75,54	752,89	73,05	739,13	72,15	751,23	76,64	741,88	73,42	735,09	76,13
Wide L-H + H-L	773,07	32,18	773,76	31,62	774,07	34,48	775,33	31,51	770,01	30,28	769,32	33,89	777,68	33,75	772,94	30,31	776,50	30,86	776,25	35,26	774,50	32,27
Z-score	-0,990		-2,627		-3,232		-1,263		-3,873		-3,161		-3,081		-4,320		-3,059		-4,220		-4,766	

Table 8. Difference between the two means (wide spread)

5.6. DISCUSSION

The first point to discuss is the realism of the simulation and, in particular, the main simplifying assumptions, that is, HF traders' preference for market orders when spread is thin, and for limit orders otherwise. This has been taken from the literature, and matches common sense. HF traders are reluctant to take too high a risk by initiating a trade when the spread is wide, preferring instead to quote limit orders. It must be noticed that when the spread is thin, aggressive orders are less risky (as chances are higher than future price swings will allow to close the position favourably) while market making still ensures profits, albeit tiny, to be made. On the contrary, when the spread is wide, market making is potentially more profitable, although profits are less likely, and aggressive trading much riskier, as closing one's position profitably becomes more of a challenge. Quite obviously, HF traders are much faster than LF traders. This statement has been relaxed in the second approach in order to verify whether or not it had an impact on the outcome.

The figures in table 4, when the speed difference between fast and slow traders was implemented, show a certain trend of both HF and LF traders to trade within their own category when the spread is, respectively, thin or wide. Therefore, it looks like the market shows some tendency to create two distinct tiers. As soon as the trading ratio approaches the one empirically found in the most technologically advanced markets, more than half of all thin spread trading goes between HF traders. This seems a clear indication of a two tier market. It must also be noticed that in the case of thin spread, all mixed trading is always lower than 1/3, a fact that seems to go the same direction as the previous statement. It is important to keep in mind the realistic, yet bold, assumption of HF traders being reluctant to trading with spread wider than one tick, which obviously leads to LF traders being relatively more active in such a scenario. At low levels of QR, more than 40% of all wide spread trading occurs between two LF traders and this figure is higher than the percentage of mixed trading. Yet, when the QR rises, the latter statement is no longer true. This means that, although LF traders still trade considerably with their similar counterparts, the main form of trading is the mixed one. The tendency of intra-community trading is present when the spread is wide but becomes progressively weaker as the ratio between HF and LF trading increases.

The situation changes significantly when the two communities experience the same latency (as shown in table 6) and the only difference between them is the reluctance of HF traders to operate aggressively when the spread is wide. The high percentage of HF-to-HF trading does no longer seem an indication of a two tier market but simply a consequence of the overall high percentage of HFT when the spread is thin. This inconclusiveness is corroborated by the intra-lane trading (i.e. the trading executed between two traders belonging to the same category, either HF or LF), which is around 50% of all trading in case of thin spread for low values of QR and that drops below 45% only for QR greater or equal to 100. The situation is even less clear when the spread is wide. The percentage of L-L trading is comparable, or even lower, than the percentage of mixed trading.

Summarising the results across quoting ratios, it can be said that:

(i) when the spread is thin, the more HF traders act and the more they tend to deal with other HF traders. The relatively high level of HF-to-LF, and vice-versa, trading must not deviate attention from the fact that quoting limit orders when the spread is thin runs a greater risk of being adversely selected. In such a scenario market taking is more prone to generate profits (in this case, mostly for the more active HF traders), although the simulation does not make this visible;

(ii) when the spread is wide, market making becomes more profitable and HF traders seem content to leave slower investors taking the initiative. It must be remembered that limit orders quoted by HF traders are frequently cancelled, leaving limit orders quoted by LF traders sitting on the book most of the time. Obviously, as HF traders increase their activity (mostly by posting limit orders), their trading activity also increases but LF-to-LF trading remains the dominant activity in such a scenario.

Overall, there is some evidence that the tendency toward market tiering is a direct consequence of the large speed difference between HF and LF traders and not of the simplifying assumptions made in the simulation.

5.7. CONCLUSION

The literature highlighted some convincing reasons for investigating the matter of a possible market tiering phenomenon more deeply and on a quantitative basis than has been done so far. In particular, the phase transition, which seems to occur in some specific cases, may lead to the market splitting into two lanes, one for the fast drivers and one for pedestrians and bicycles. This suggested to implement a simulation for verifying the soundness of the 2-tier hypothesis. After analysing the outcome of the simulations, some tendency appears but it is not completely clear if it can be considered a definite market behaviour. Market tiering seems more evident among HF traders when the spread is thin than among LF traders when it is wide, and overall it is possible to claim a certain evidence (stronger with thin spread and weaker with wide spread) that some degree of market tiering occurs but not sufficient, since the evidence obtained is not strong and unambiguous enough, to recommend actions from the regulators based on this evidence only. Further research is necessary before any definitive result, in one sense or the other, could be claimed. As found by Cartlidge and Cliff (2012) and by Johnson et al. (2013), there seems to be no step change, or phase transition, but rather a smooth trend towards market tiering as more HFT takes place. A possible way out of the limit outlined above, would be accessing and investigating in depth confidential data owned by the exchanges, which permit to identify the traders' identity. Such data is not publicly available as the agreements between investors and the exchanges usually dictate so in order to protect the privacy and the trading strategies of the participants. If the trading strategies were public, it would be too easy for competitors to take advantage of them, by foreseeing and front-running such strategies. However, in some cases, researchers have been granted access to such information under Non-Disclosure Agreements, examples being Kirilenko et al. (2011), Menkveld and Yueshen (2016) and, of course, the CFTC-SEC (2010a, 2010b) reports. Being able to categorise the traders would allow to better understand the trading distribution according to the four classes, HF-HF, HF-LF, LF-HF, and LF-LF (where, as usual, the first party represents the aggressive side), split by spread width. Another dimension useful for the research is the HFT ratio and how tiering behaves as the ratio varies. Achieving the required 40% to 60% trading ratio by using a simulation is relatively straightforward. In the case of the simulation above, the independent variable to play with was the quoting ratio, QR. By slowly changing the QR, it was possible to correspondingly vary the trading ratio. With real data this may prove not so easy; it may require digging deep into the data until a time window displaying the target trading ratio is found. A trial-and-error approach therefore seems to be necessary.

6. HIGH-FREQUENCY TRADING AND THE EFFICIENT MARKET HYPOTHESIS

6.1. INTRODUCTION

The original concept of efficient market dates back to the XIX century, when the French broker Jules Regnault hypothesised the random behaviour of the stock markets, later supported by his fellow countryman Louis Bachelier, at the beginning of the 1900. Recognition of the random walk was further developed into the so-called Efficient Market Hypothesis by Eugene Fama in the late Sixties. Among the advocates of the non-validity of the Hypothesis, there are the supporters of Technical Analysis (TA). In the Western World, practitioners have used TA, although to different degrees of depth, for nearly one and a half century, since Charles Dow first published some simple stock indexes on the "Customer's Afternoon Newsletter", although according to Kirkpatrick and Dahlquist (2007) the origin of TA to some extent can be dated as back as the Sixteenth century in Japan. Despite its history, the academic world has often dismissed TA, at best, as some sort of self-fulfilling prophecy.

In the following I shall contrast the impact HFT has on arbitrage against the classical theories, which obviously did not take ultra-fast trading into account. Then, some arguments in favour and against the existence of arbitrage opportunities will be illustrated, before providing quantitative arguments to support the thesis of significant impact of HFT on market efficiency. In order to do so, it has been developed a computer-based simulation of a two-market environment in which a small number of HF traders interact with a much larger number of slow traders face to a significant price difference for the same security in one venue with respect to the other. Such difference gives rise to an arbitrage opportunity, which all the traders will rush to profit from. The results of the simulation will be stored onto a database and analysed and compared to the findings of the current literature.

6.2. IMPACT OF HIGH-FREQUENCY TRADING ON ARBITRAGE

6.2.1. INTRODUCTION

The most common opinions among academics are that HFT is, in general, beneficial as far as market efficiency is concerned. Aitken et al. (2012) provide evidence of improved efficiency in the wake of increases in HFT, even though the same paper admits that "despite being in the spotlight for some time, our understanding of high-frequency trading and its implications for market quality are at best, moderately informed" (ibid. p.3). However, the authors take the cost of trading as a proxy for efficiency, although the term 'efficiency' in this context is more akin to the meaning of 'efficient allocation of resources' rather than to the one used by Nobel laureate Eugene Fama in his original explanation of the Efficiency Market Hypothesis (no consistent risk-free abnormal profit). Aitken et al. (2012) also find that most academic papers agree on a "predominantly positive overall impact" (ibid. p.8) of HFT on market efficiency. However, they also quote a study by McInish and Upson (2012), who developed a theoretical framework modelled on the US equity market that demonstrates how, under certain conditions, HF traders can profit from their superior knowledge of the current state of the market, thanks to both faster data analysis capabilities and quicker access to the order books, versus slower competitors. Price discovery is another proxy often used for market efficiency, as it is closely linked to arbitraging. If the price is not aligned with the true value of a security, then arbitrage opportunities arise, and HFT operators are in the best position for assessing and exploiting them, wiping out such opportunities, as expected by the EMH. The study by Aitken et al. (2012) demonstrates this result by computing marking-the-close, information leakage, effective spreads, and cancel-to-trade statistics on the LSE and Euronext-Paris data between 2001 and 2011. Also Easley, Lopez de Prado and O'Hara (2012) state that "[o]ver short intervals, prices are not the random walks so beloved by the efficient market hypothesis, but can instead be predictable artifacts of the market microstructure" (ibid. p.20)), where 'short interval' (in 2012) very likely indicates the realm of HFT. All the considerations described above highlight the advantage of HF traders with respect to other market participants thanks to the formers' ability to exploit their own speed advantage coupled with transitory inefficiencies of the markets.

6.2.2. THE IMPORTANCE OF TIMING

However, there is another aspect that highlights the existence of market inefficiencies even in presence of all EMH pre-conditions. Let us assume that the markets are efficient in the semi-strong form. It means that, as shown by Fama (1965) and Fama (1970), as soon as an arbitrage opportunity arises, it will be instantaneously discounted by the market. As discussed above, the word instantaneously is relative to the technological environment, the common requirement being that arbitrage opportunities are incorporated quickly enough into prices by the market. But an efficient market is made up of a large number of participants - and it is the entire community of the market participants that adjust prices according to the newly created situation. If, instead, the price movements were not caused by new hard facts but only by 'noise', then again arbitraging would bring them back to their true value. There are several types of noise in markets. Kirkpatrick and Dahlquist (2007) suggest four ways in which bad information may temporarily affect prices: (i) new information may be inaccurate; (ii) the source may intentionally disseminating false information for disguising the market; (iii) dissemination latency may be not negligible; (iv) interpretation latency is not the same for all recipients. Aldridge (2010) also reminds about information leakage before official announcements. In all the cases mentioned above (and in others), prices will sooner rather than later tend to re-adjust to true value, even in absence of new information, by the very same definition of noise. So, even according to the EMH, some market operators (called arbitrageurs) actually do make abnormal profits when they re-adjust the price exploiting an arbitrage opportunity. Malkiel (2003) puts it stating that "any truly repetitive and exploitable pattern that can be discovered in the stock market and can be arbitraged away will self-destruct" (ibid. p.72) and Wilson and Marashdeh (2007) recognise that arbitrage is a sort of short-term inefficiency that ensures long-term efficiency. A paradox seems to arise at this point: for the markets to be efficient arbitrageurs must exist to bring arbitrage opportunities down by exploiting them, but if they do so, abnormal risk-less profit would arise, showing the markets are inefficient. The Efficient Market Hypothesis had to deal with arbitrageurs somehow and, if their existence could not be denied, the

Hypothesis had at least to show some special characteristics of theirs. The way it solves the paradox was by stating that nobody is able to make abnormal profits 'in a consistent way', by exploiting arbitrage opportunities because all rational investors (that is, all investors, except the most naïve or casual ones) will compete to do so. In an efficient market abnormal profits are therefore split between all (or at least between very many) operators, which sometimes reap the abnormal profit through arbitrage and sometimes, finding themselves on the wrong side of the arbitrage, suffer a loss [Fama (1965)]. With a sufficiently large number of observations, the abnormal profit for each arbitrageur averages to zero or near zero [Van Horne (1995)]. At the end of the day (or the month, or the year) the cumulative abnormal profits and losses for each operator ought to tend balancing each other out, reaffirming the market efficiency principle. Yet, there is a big difference between never making any abnormal profit and averagely making no abnormal profit. Again, if one participant, or a small group (relative to the total number) of participants, like HF traders, is consistently able to arrive first to exploiting arbitrage opportunities, then the Hypothesis does no longer hold for them, although it may still hold well, on average, for the market as a whole, or the whole arbitrage community. But this would just be a more sophisticated version of the classical joke defining statistics as averaging between the one who eats two chickens and another who eats none, resulting in everyone eating one chicken. Shostak (1997) argues that "even if we are to accept that modern technology enables all market participants equal access to news, there is still the issue of news interpretation" (ibid. p.29), which HF traders are likely to exploit before non-HF traders can. But well before the issue of news interpretation, HFT is an issue of modern technology that does not enable all participants equal access to the news, because ultra-fast access is not equal for all - quite the opposite, it is the main reason for technology-led market inefficiency. Obviously, the inefficiency only persists as long as the fast arbitrageurs are a small minority of the investor community. If they become the majority (or ideally, the totality) of the market participants, then the Efficient Market Hypothesis would hold again, as the average abnormal profit for each participant (and not just on average for the totality of them) over the long-term would revert to the mean expected by the EMH: zero.

6.2.3. ARBITRAGE AND TRANSACTION COSTS

As first devised by Fama (1970), transaction costs are a tool leading toward market efficiency. If small discrepancies are found, they may not be exploitable because of transaction costs, which would make the trade unprofitable. Under these conditions, the discrepancy is immaterial and does not exist from any practical point of view. This is true for all market participants and for HF traders too, but some considerations are to be taken into account.

Impact analysis of the Markets in Financial Instruments Directive in some sense leads Aitken et al. (2012) to turn the table by stating that the regulatory changes are the cause of both implicit and explicit trading cost falling, from which HFT benefited, rather than the latter being the cause of the former. Sornette et al. (2011), despite generally criticising HFT practice, recognise that HFT does improve liquidity, and that higher liquidity and higher volumes traded (usually linked to HFT activity - the same paper reports HFT making between 60 and 70% of equity trading volume) tend to lower transaction costs. According to Jain (2005), computerized trading systems lower spreads, fees, brokerage, and commission costs but Friederich and Payne (2011) suggest that the originator of such market efficiency improvements may not be HFT activity. Brogaard et al. (2014) go so far as to argue that HFT may also be the cause of an increase in transaction costs. They also highlight that most academic literature supporting the view according to which HFT lowers transaction costs, does only take into account execution costs, leaving to discuss other components of transaction costs, as commissions and technology costs.

6.2.4. IS ARBITRAGE A REAL ISSUE?

After having shown that HF traders can exploit arbitrage opportunities and make consistent abnormal profits, it is important to understand whether or not arbitrage opportunities are a real thing, and how often they occur in real life.

In one of his papers on the EMH, Fama (1970) deals with this issue, concluding that departures from the independent price assumption, whereas they may deny the random walk model, are not incompatible with the Efficient Market Hypothesis. The matter is settled by recognising a role of arbitrage in absorbing minor market inefficiencies - and avoiding them to accumulate up to an intolerable level. The beneficial role of arbitrage is also appreciated by Shleifer and Vishny (1997), because of its effect in bringing prices back to fundamental values and to keep markets efficient. However, the critical point of involving the investor community at large, and having them to share the profits brought in by arbitraging cannot hold the analysis carried out by Shleifer and Vishny (1997): "the millions of little traders are typically not the ones who have the knowledge and information to engage in arbitrage. More commonly, arbitrage is conducted by relatively few professional, highly specialized investors who combine their knowledge with resources of outside investors to take large positions" (ibid. p.2).

6.2.4.1. News Arbitrage

When news reach the market, it sometimes has unclear consequences but at some other times consequences are crystal clear. With automatic news feed interpretation, those who enjoy highest speed can exploit some unambiguous news to their own advantage. According to the classic Efficient Market Hypothesis, markets would instantaneously discount newly arrived information where the focus is on the word 'instantaneously', which may carry different quantitative meaning for different market participants. The casual trader may become aware of the information during the evening news television programme, whereas a professional trader may know it via its Bloomberg screen, and a HF trader equipped with automatic news feed interpretation may, by then, have already closed all the news-related transactions. This is undeniably an arbitrage opportunity.

6.2.4.2. Latency Arbitrage

In the HFT world, the focus is mostly on latency - and according to Arnuk and Saluzzi (2009), arbitrage is no exception: "We believe Latency Arbitrage is more than a simple case of technological evolution, but raises serious questions about the fairness and equal access of US equity markets" (ibid. p.1). They compare HFT to someone having access to the relevant news contained in the Wall Street Journal five microseconds into the future, and they report that thanks to this information, HFTs are able to achieve "(almost) risk free arbitrage opportunities" (ibid. p.2). The example they show is instructive.

"1. The book for stock ABC shows $25.53 bid / $25.54 ask.

2. Due to Latency Arbitrage, a HFT computer realises that there is an incoming order that in a fraction of a second (but still a long time for HFT's standards) will move the NBBO quote higher, to $25.54 bid/offered at $25.56. [How and why a HFT computer may get this information is explained in detail in the following section.]

3. The HF trader hurries up, scraping dark and visible pools, buying all available ABC shares at $25.54 and cheaper.

4. When eventually the order posted by the institutional algorithm hits the server exchange, it cannot have its bid order at $25.54 executed as there is no longer stock available at this price and the market moves up to $25.54 bid / $25.56 ask as anticipated by the HFT after inspecting the order books.

5. The HF trader then turns around and offers ABC at $25.55 or $25.56.

6. Because it is following a volume driven formula, the institutional algorithm is forced to buy available shares from the HFT at $25.55 or $25.56.

7. The HFT makes $0.01-$0.02 per share at the expense of the institutional investor." (ibid. p.2)

With this tiny profit per share and per execution, Arnuk and Saluzzi (2009) calculate a daily profit ranging between $6 and $12 million, which multiplied by 250 trading days per year translates into $1.5-$3 billion profit generated at the expenses of retail and institutional investors. Obviously, there is no guarantee that on every trading day an institutional investor will launch a large buying or selling programme but the idea is clear.

According to a HFT software firm Advent's white paper, Advent (2012), "[i]n most instances, high-frequency traders don't bet on the value of a company, currency or commodity - or the future outlook - but are simply looking to arbitrage price discrepancies in securities that are trading simultaneously on different exchanges or trading platforms" (ibid. p.7).

6.2.4.3. Arbitrage Evidence

Practical evidence of how latency arbitrage works is described in 'Flash Boys' [Lewis (2014a)]. Everything started when in early 2007 the trader team led by Brad Katsuyama realised that when they tried to complete a trade it became nearly impossible to do it properly, as the shares on offer suddenly vanished. Initially they thought of a computer or connection problem but a deeper investigation showed that the market seemed to specifically respond to their action by removing liquidity as and when they were about to take it. As it turned out, it was not their exclusive problem - it was a market-wide issue. Katsuyama's team launched a deep study of the problem on 13 stock exchanges scattered over four different sites run by NYSE, NASDAQ, BATS and Direct Edge (located in Weehawken, Secaucus, Carteret, and Mahwah, all in New Jersey, USA). They discovered that their orders were properly executed when sent to an individual exchange but did not when launched to all of them at the same time. One exception was noticed: no matter how many exchanges were their orders sent to, they always achieved 100% execution from the exchange located in Weehawken. Further investigation revealed that orders to Weehawken were always 100% successful because that exchange was located nearest to their premises. Further away exchanges did show phantom liquidity. Indeed, when a modification to their trading platform arranged to have orders to reach all the exchanges at the same time (instead of the nearest one first and the others in sequence according to distance), it turned out they achieved 100% successful execution. The issue might be simplified by modelling a trader T and two exchanges, E1 and E2, located at different distances from T. Let us suppose trader T reads the price P for stock S, which it considers appealing and then it sends a relatively large market order for a quantity Q of shares to the market, which is routed to E1. Yet, E1, the nearest exchange to T, cannot trade all Q shares as requested by T, because it can only provide for nQ of them, with $0<n<1$. So, E1 accepts the order for nQ shares, executes it and routes the order for the remaining (1-n)Q shares to E2. The discovery made by

Katsuyama and his team, as described in Lewis (2014a) is that the second order would never be executed at the originally intended price because the shares on the book would be bought (or sold) by faster arbitrageurs before T's order arrived there.

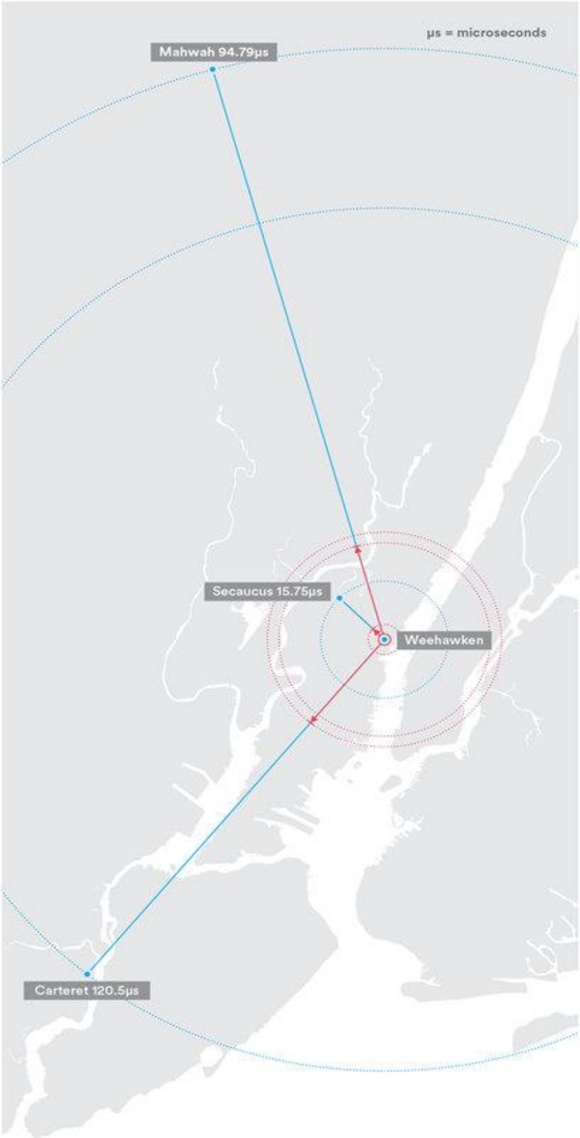


Figure 9. In the microseconds it takes a high-frequency trader — depicted in blue — to reach the various stock exchanges housed in these New Jersey towns, the conventional trader's order, theoretically, makes it only as far as the red line. Reproduced from Lewis (2014b)

Indeed, a HF trader, noticing the order filled on E1 at price P, would immediately (exploiting its superior technology and faster connections) front run T's order to E2, trading all shares at the current price P and quoting a limit order at price P +Δp (with Δp greater or less than zero according to whether T intended to buy or to sell, but still better than the next limit order for the same stock

S). When T's market order eventually arrives at E2 the price it is being executed at is worse than originally intended, to the benefit of the HF trader, somehow 'frontrunning' it. What was happening to Katsuyama's team (and many others) was exactly that. Confirmation of this scenario comes from Cohen and Szpruch (2012), who build a mathematical model representing two investors, a HF and a LF trader, operating on the same security market. Their little surprising conclusion is that "[i]f the fast investor can front-run the slower investor [the situation Katsuyama's team find themselves in], we show that this allows the fast trader to obtain risk free profit" (ibid. p.211). This practice may be a legal form of frontrunning, but still leading to abnormal, consistent, risk-free profit. It is a form of arbitrage and as such it conflicts with the EMH.

Practitioners are generally rather suspicious about HFT. The ZeroHedge blog by Tyler Durden is a very popular website among professionals; it publishes articles and receives lots of comments on a wide range of financial matters and not surprisingly HFT is a hotly debated issue. About market efficiency, Durden (2014) states "High frequency traders use ultra-high speed connections with trading venues and sophisticated trading algorithms to exploit inefficiencies created by the new market structure and to identify patterns in 3rd parties' trading that they can use to their own advantage". A critical piece of information is the investment pattern followed by algorithms (usually more predictable than humans). "Information is leaked when electronic algorithms reveal patterns in their trading activity. These patterns can be detected by HFTs who then make trades that profit from them" [Durden (2014)], leading to the sharp conclusion "HFT is legal frontrunning".

6.2.5. DO ARBITRAGE OPPORTUNITIES EXIST IN THE REAL WORLD?

Fama (1965) acknowledged arbitrage opportunities in his theory - stating that even if they occur, the market forces would wipe them out immediately, leading back to efficiency. If the only evidence of arbitrage was the one described by Lewis (2014a), it could be considered an exception. In practice several authors investigated their existence. The Law of One Price (LOP) in international financial markets is investigated by Akram, Rime and Sarno (2009), who find that it "holds on average, but numerous economically significant violations of the LOP arise. The duration of these violations is high enough to make it worthwhile searching for one-way arbitrage opportunities" (ibid. p.1741). Interestingly, they find that "such opportunities decline with the pace of the market" (ibid. p.1741), implicitly suggesting a dependency from speed which their study does not investigate further to its consequences. Focusing on the American Depository Receipts market, Wahab, Lashgari and Cohn (1993) combine long and short positions forming arbitrage portfolios. By undertaking two perturbations, namely minimum-variance and an equally-weighted combination, their research leads to rejecting the null hypothesis of no arbitrage. This result differs from the one previously found on the LOP by Kato, Limm and Schallheim (1990), according to which "no obvious arbitrage opportunities exist between the international capital markets" (ibid. p.73), although they acknowledge that "[t]he two returns are not near to being perfectly correlated" (ibid. p.73). In the same year Brenner, Subrahmanyam and Uno (1990) found that the Japanese futures contracts showed persistent departures from their fair price, opening the way to potential arbitrage opportunities. In more recent times, controversial support to the Efficient Market Hypothesis is provided by Hens, Herings and Predtetchinskii (2006), who recognise the existence of arbitrage opportunities but state the difficulty to make use of them on the basis that in order "to exploit all existing arbitrage opportunities, traders should pay attention to all financial markets simultaneously" (ibid. p.556), something very hard to even remotely approach in 2006, at the dawn of HFT practice. Cross-market arbitrage opportunities are investigated by Carter (1989), who compares long-term government bonds futures between the thinly liquid Toronto Commodity Exchange and the more liquid Chicago Board of Trade, under the hypothesis that, because of its higher price variance, the former could be profitably arbitraged against the latter. Data refer to the period September 1982 to June 1985 and the outcome leaves no doubts: "the results show there are arbitrage opportunities in the bond futures market" (ibid. p.352). A similar conclusion is reached by Kolb, Gay and Jordan (1982): "the results of this study do not justify the conclusion that the market is efficient" (ibid. p.228).

Market crosses (bid price higher than ask price) across different market venues generate arbitrage opportunities according to Garvey and Murphy (2006), who also find statistically significant profitability for "institutional traders who act fast and pay little in trading costs" (ibid. p.57). Ghadhab and Hellara (2015) find that price deviations for cross-listed stocks on US and European exchanges show that markets are not efficient. The last study in this list (from a much longer one) is Gagnon and Karolyi (2010). The two scholars select a sample of more than 500 companies from 35 countries whose cross-listed shares are traded on the US as well as on local exchanges and compare the prices of the stocks over different exchanges on a currency-adjusted basis. The result is a relatively small average deviation from price parity, 4.9 basis points, which is often higher than most transaction costs. This alone would make the arbitrage opportunities potentially worth exploiting. Moreover, the authors also find a more pronounced 1.4% fluctuation above the average for the typical stock pair. Danielsson (2013) puts it rather clearly: "the bulk of the activity in FX [Foreign eXchange] markets is due to speculation, the buying or selling of currencies solely to profit from anticipated changes in exchange rates, by means of high frequency trading (HFT), exploiting arbitrage between currencies (triangular arbitrage) and prices on different trading platforms" (ibid. p.195). Although there are also studies that find, according to the theory, that arbitrage opportunities are non-existent after taking transaction costs into considerations, the view of those claiming inefficiencies cannot be dismissed and in recent times it seems the dominant one among scholars.

6.3. AN ARBITRAGE SIMULATION

6.3.1. INTRODUCTION

In order to provide quantitative evidence to what discussed above, this paragraph presents a computer-based simulation to test the null hypothesis that it is impossible to claim consistent abnormal risk-free profits out of arbitrage against the alternative hypothesis that it is possible. Yet, before going ahead with the simulation, it was important to verify whether arbitrage opportunities do actually occur in the real world, as it has been done in section 6.2.4. The next section explains the methodology used, then the following section (6.3.3.) describes the details of the simulation and section 6.3.4 presents the results. The final section (6.3.5) discusses the results, with particular focus on the soundness of the methodology, on the impact on non-HF investors, and on whether the outcome should alert market regulators.

6.3.2. METHODOLOGY

The simulation is made up of two markets trading the same set of securities. At each cycle a discrepancy arises in one of the markets by forcing in one randomly-chosen security a price change that allows an arbitrage opportunity. All participants immediately notice that price difference but, since the purpose of the simulation is to provide evidence of the principle, not working out the profit made through arbitrage, for sake of simplicity and without loss of generality, it was set that one and only one trader is able to grasp the opportunity and that just one security is traded at every cycle. One randomly chosen trader buys the security in the market where the price is lower and sells it in the other market. The only difference between traders is their speed: HF traders, thanks to their higher speed, have higher chances to be randomly selected and so to exploit the price difference.

6.3.3. DESCRIPTION OF THE SIMULATION

The user-defined parameters read at the beginning of each run are now described. The number in brackets is the values used for the corresponding parameter in the simulation.

- Number of traditional traders (11,859)
- Number of HF traders (15)
- Number of securities traded in both markets (500, e.g. the securities in the S&P 500 index)
- Number of cycles to run the simulation for (100)
- Maximum allowed price variation in cash units (1)
- Trading ratio between HF and traditional traders (381)
- Transaction costs, including spread between the bid and ask price and fees (3.5 basis points)

The reason for having several repetitions of the simulation is to provide a suitably high number of data that allows using the standard deviation of the sample as replacement of the standard deviation for the entire population (that is, making use of the large sample Z-statistics). The reason for having several iterations (100 cycles in this case) within the simulation is to mimic a suitable number of arbitrage opportunities occurred within the observation period. A quantitative estimation of transaction costs depends on various factors, including the venue, the fee structure, the bid-ask spread (which on its turn depends on market conditions), and possibly others. I have used 3.5 basis points as a rough estimate, bearing in mind that, according to many academic studies, higher HFT activity contributed to lowering transaction costs but also that too low an estimate of transaction costs would falsely increase arbitrage profits. Subsequent launches of the simulation increased commissions to investigate the quantitative level of trading costs at which the EMH would hold. For technical reasons, every computer-based random number generation needs a lower and an upper limit. For initialising the security price, I set those limits to 20 and 80 cash units (being GBP, USD, EUR or any others, or even fractions of these currencies). Whereas this choice is arbitrary, it is always possible to write arbitraging algorithms that act only upon securities whose price change falls within a predetermined range, so my choice does not drive the result in a way or another. Arbitraging traders are set to hold no securities throughout the simulation period, to sterilise their final account of any market gain or loss; the difference between their final and initial account only entails cash variations due to exploitation of arbitrage opportunities. The numbers of HF and non-HF traders are taken from CFTC-SEC (2010b), where at page 29 it displays summary statistics whence it is possible to extract the number of HF and non-HF (split among Intermediary, Buyer, Seller, Opportunistic and Noise) traders active in the E-mini contracts market on May 6, 2010 as well on the previous three days. I took the figures from the latter statistics in order to avoid any reference to an exceptional situation as the day of the Flash Crash. The same statistics report the percentage of trades carried out by the different categories of traders and it is therefore possible to work out the ratio between the HF versus non-HF trades on those three days. According to CFTC-SEC (2010b) the 15 HF traders carried out 32.56% of all trades, whereas the other categories (11,859 traders altogether) shared the remaining 67.44% trades. Therefore, the average HF trader took 2.17% of all trades whereas the average traditional trader only closed less than 0.0057% of the trades. The ratio between the two figures yields 381.7, which I took freedom to round down to 381 HF trades for every 1 non-HF trade. The pseudo-code with a more detailed explanation of the algorithm can be found in Appendix C.

6.3.4. RESULTS

The run of the algorithm yields an average profit for each HF trader of 3.17 cash units (against the null hypothesis of zero or near zero, as expected by the EMH) and a standard deviation of 0.22 cash units. Bearing in mind the considerations above about transaction costs, the simulation was also launched with a figure ten times higher (35 basis points instead of 3.5), which still yielded an average profit for HF traders of 1.97 cash units and a standard deviation of 0.19 (table 9). The result is extraordinarily strong for both levels of transaction costs, at virtually any significance level, against the null hypothesis of no consistent risk-free gain. Therefore this simulation leads to reject the null hypothesis - and to reject the Efficient Market Hypothesis in presence of HF traders with it.

Costs (bp)		Profit of HF traders	Profit of other traders	Skewness	Kurtosis	Bowman-Shelton
3.5	Avg	3.17399	0.0000132	0.34615	3.01669	1.99819
	StDev	0.22129	0.0000293			
	Z-score	143.42825	4,50417			
35	Avg	1.97287	0.0000055	0.13150	2.34124	2.09639
	StDev	0.19136	0.0000149			
	Z-score	103.09662	3.70846			

Table 9. Average and stddev profit for HF and non-HF traders

At the contrary, traditional traders are not able to make consistent abnormal profits, as the profit they make (1.3x10^{-5} cash units in case of 3.5 bp costs and 5.5x10^{-6} in case of 35 bp costs) is, under any practical respect, negligible. This result is compatible with Fama (1970), who claims that "it is possible to devise trading schemes based on very short-term (preferably intra-day but at most daily) price swings that will on average outperform buy-and-hold. The average profits on individual transactions from such schemes are miniscule, but they generate transactions so frequently that over longer periods and ignoring commissions they outperform buy-and-hold by a substantial margin. […] But when one takes account of even the minimum trading costs that would be generated by small filters, their advantage over buy-and-hold disappears" (ibid. pp.395-396). This result also confirms Van Horne (1995), who finds profit from arbitraging averaging to 'near zero', which is exactly the outcome of the simulation for LF traders. It must be said that, should the number of HF traders rise to reach the majority of the participants, it is reasonable to expect arbitrage profits to be shared among all of them, so reducing the numeric amount to the one obtained by LH traders. In such a homogeneous scenario the Efficient Market Hypothesis would hold again.

In order to correctly apply hypothesis testing, the data distribution needs to be normal. The Bowman-Shelton test verifies normality of data distribution. It uses a formula based on skewness and kurtosis [Newbold (1995), p.413]:

$$B = n\,[S^2 / 6 + (K - 3)^2 / 24]$$

where n is the number of observed data, S is the values of skewness and K the value of kurtosis.

The procedure is to calculate the above statistics and to reject the null hypothesis if B exceeds the value corresponding to the sample size in the 'significance points of the Bowman-Shelton statistic' [Newbold (1995), p.414], reproduced in table 10.

Sample size n	**10% point**	**5% point**
20	2.13	3.26
30	2.49	3.71
40	2.70	3.99
50	2.90	4.26

75	3.09	4.27
100	3.14	4.29
125	3.31	4.34
150	3.43	4.39
200	3.48	4.43
250	3.54	4.51
300	3.68	4.60
400	3.76	4.74
500	3.91	4.82
800	4.32	5.46
∞	4.61	5.99

Table 10. Significance points of the Bowman-Shelton statistic [reproduced from Newbold (1995)]

The hypothesis testing technique is applicable since the data obtained by launching the simulation display a normal distribution, with a skewness of 0.346 and a kurtosis of 3.017, which yield a value for the Bowman-Shelton's normality distribution test of 1.998 (for the case of costs equal of 3.5 basis points). Also, when the costs are equal to 35 basis points, the values of skewness, kurtosis and Bowman-Shelton test are 0.131, 2.341, and 2.096, respectively.

Since the values obtained for the Bowman-Shelton statistics do not exceed the tabulated values in table 10 for sample size equal to 100, the null hypothesis of normality distribution of the data cannot be rejected at either 10% or 5% significance point level. This result confirms the normality of both distributions and therefore the applicability of the Z-statistics for large samples [Trivedi (2016)].

A similar result, albeit with a smaller average, but still abnormal, profit was obtained with costs 20 times higher than the base case (70 basis points). An outcome as expected by the Efficient Market Hypothesis (null or negligible profit for HF traders) was only achieved with costs somewhere between 80 and 90 basis points (table 11). Again, non-HF traders are not able to display any significant abnormal profit, confirming the outcome expected by the Efficient Market Hypothesis [Fama (1970)]: 2.9x10^{-6} cash units, -1.1x10^{-6}, and -1.8x10^{-6}, respectively for costs equal to 70, 80 and 90 bp. The values of skewness and kurtosis are such that the Bowman-Shelton conditions for normality of distribution, following the same procedure as above, are met.

6.3.5. DISCUSSION

The result achieved raises three types of considerations or questions:

(i) the significance of the distorting role of HFT to market efficiency in the simulation is so strong that it is legitimate to suspect that the outcome was implicit in the assumptions;

(ii) should other investors worry about the impact HFT seems to have on market efficiency?; and

(iii) is the result something that should attract the regulators' attention or is it just another chapter of the old story according to which smart investors have always made more money than (and to the expenses of) the dumb ones?

Costs (bp)		Profit of HF traders	Profit of other traders	Skewness	Kurtosis	Bowman-Shelton
70	Avg	0.52484	0.0000029	- 0.11724	2.60419	0.88186
	StDev	0.18547	0.0000109			
	Z-score	28.29812	2.62640			
80	Avg	0.17185	- 0.0000011	- 0.10622	2.70149	0.55934
	StDev	0.17533	0.0000102			
	Z-score	9.80182	- 1.07755			
90	Avg	- 0.23041	- 0.0000018	0.05909	3.25511	0.32937
	StDev	0.20771	0.0000114			
	Z-score	- 11.09290	- 1.53828			

Table 11. Average and stddev profit for HF and non-HF traders

6.3.5.1. Is the Simulation Just Proving Its Assumptions?

The result is so strong that suspicions of a pre-determined test whose outcome is implicit in its starting assumptions cannot be easily dismissed. Yet, several authors, Aldridge (2010) among others, share the opinion that "[w]hoever detects the mispricing and gets his order posted on the exchange first is likely to generate the most profit" (ibid. p.245). According to Jarrow and Protter (2012) "with high frequency traders, we can show that there exist no arbitrage opportunities for ordinary traders" (ibid. p.2), which implicitly means that all arbitrage opportunities would be exploited by non-ordinary (i.e. HF) traders. Similarly, Anderson (2016) finds evidence that statistical arbitrage is feasible and it is ‘plausible’ that a significant part of it is due to HFT: "compelling evidence of a consistent price lead and lag relationship between individual stocks at the high-frequency level. This shows that statistical arbitrage, and therefore high-frequency trading, is entirely possible as far back as 1999" (ibid. p.216). Another confirmation of how concrete HFT arbitrage opportunities are, is given by the success of the Alternative Trading System IEX, launched by Katsuyama and his team [Lewis (2014a)] with the purpose of 'speed bumping'. The engine introduces a 350ms automatic delay, which slows down ultra-fast trading, levelling the playground for all. According to Buchanan (2015), "IEX has already attracted about 1% of stock-trading volume in the United States" (ibid. p. 163). Budish, Cramton and Shim (2015) use millisecond-level direct-feed data to analyse stylised facts about how markets work in a HFT scenario. They find at that time horizon: "(i) correlations completely break down; which (ii) leads to obvious mechanical arbitrage opportunities; and (iii) competition has not affected the size or frequency of the arbitrage opportunities, it has only raised the bar for how fast one has to be to capture them" (ibid. p. 1548). Even sharper is the finding of Scholtus, van Dijk and Fijns (2014), who view speed as "crucially important for high-frequency trading strategies based on U.S. macroeconomic news releases. Using order-level data on the highly liquid S&P 500 ETF traded on NASDAQ from January 6, 2009 to December 12, 2011, we find that a delay of 300ms or more significantly reduces returns of news-based trading strategies" (ibid. p. 89). Obviously, the best arbitrage opportunities are picked up by the fast traders, as also pointed out by Pirrong (2014): "Competitive markets are fragmented. [...] Fragmentation inherently creates arbitrage opportunities, and arbitrage profits go to the swift" (ibid. p.14). Finally, Vella and Ng (2016), while (albeit diplomatically) dismissing the validity of the Efficient Market Hypothesis in the new environment, highlight "the need for new theories in support for high frequency financial phenomena during which the human traders lose the ability to react in real time" (ibid. p. 84).

Having showed that many authors agree in principle with the existence of arbitrage opportunities in a HFT trading environment, it remains to argue that the simulation presented in this chapter does not artificially produce abnormal profits. The initialisation parameters of the simulation have been discussed above and they have all been taken from the real world. The only exception is the number of arbitrage opportunities (called iterations in the algorithm), which has been arbitrarily set to 100. The reason for using a relatively large number is to smooth out the randomness of the individual arbitrage opportunity. The number of iterations balances out the utterly unrealistic simplification of trading only one security per each arbitrage opportunity. The general idea is rather obvious: if a small group of traders has a definitive speed advantage over the large remaining majority, it is intuitive that as soon as an arbitrage opportunity arises, it shall likely be picked up by one of the fast traders. Given large amount of literature affirming HFT advantage in exploiting arbitrage opportunities, the results of this simulation may sound intuitive. Yet, in academic research even intuitive results are best to be proved quantitatively. A question may indeed arise about what will happen when the number of HF traders increase further. In the simulation, all HF traders have the same probability to exploit the opportunity and if their number increases, they will all enjoy a lesser share of the arbitrage profits. Stretching this argument to its limit, in a purely (or mostly) HFT world, arbitrage opportunities would be shared by all the participants and profits split between all of them - and the EMH would likely hold again. At this stage I would like to highlight that the EMH was a great achievement in finance and its author deservedly was awarded the Nobel Prize. The fact that later technological innovations modified the market environment to a degree absolutely not foreseeable at the time the Hypothesis was first developed, does not reduce the importance it had, and will possibly have again in a future HFT-dominated environment.

6.3.5.2. Should the Result Worry Other Investors?

There is widespread worry among (non-HFT) practitioners about HFT practices but, as seen above, academics are much more cautious. The latter opinion would obviously be of little help should slow practitioners get convinced they are playing an unfair game and thus decide to withdraw from the financial markets altogether. However, there is no sign of such a trend. Financial markets are busy as ever and poor results are usually interpreted as long-wave effect of the 2008 crisis, or other reasons, rather than caused by HFT activities.

It must be stressed the fact that this research did not purposely take into account any market manipulation factor. The purpose here is to evaluate the impact of clean, straight and transparent high-frequency trading practices have on market efficiency and not whether, as suggested by some practitioners and academics alike, HFT is suitable to be used for market manipulation purposes. HFT-led manipulative practices may well be the scope of further research. Lewis (2014a) lucidly illustrates some of the non-manipulative (as well as some manipulative) effects of HFT and in particular the slow-market arbitrage activity described there is essentially the one simulated in this paragraph.

On one side, it is true that any successful innovation worries those who cannot replicate or satisfactorily tackle it. On the other side, this situation has occurred countless times in history – and financial market history is no exception. Yet, financial markets continued to survive and to provide their services to the world economy.

6.3.5.3. Should the Result Alert Regulators?

The role of regulators is obviously different from practitioners' but everybody's opinions must be (and usually are) taken into account ("*Es gibt noch Richter in Berlin*"). The playing field must not only be perceived as even, it must be even under any respect - and it is beyond doubt that HFT practices do pose a question to regulators. On the other side, the search for speed, and in general the efforts to ensure a more favourable position with respect to others, is a recurring theme in financial markets. The floor traders nearer to the broker did actually enjoy a favourable position. Remote traders that used telegraph instead of manually dispatched orders also enjoyed a favourable position, and so did telephone users with respect to the telegraph's. Speed is often a definite advantage, but is this a valid criterion for requesting a stricter regulation of HFT practices? In many cases, some kind of advantage is accepted. Nobody would accuse a trader firm of unfair advantage just because of better quality of its software, nor because of the terabytes of its disk or the gigaflops allowed by its computer's CPU. At this point it is interesting to notice that speed of CPUs seems to be tolerated by practitioners and regulators alike, whereas speed of networks apparently rises concerns. The former advantages are rather regarded as assets to replicate than as swindles to blame, as happens to the latter. Yet, networking speed and co-location present different characteristics from other computer-related innovations. Software, CPU and disks would not turn efficient markets into inefficient ones. No matter how smart software is, it will never nullify trading risk. Software may help traders to take more informed decisions, to diversify risk in a more rational manner, to replicate more closely past successful strategies – but the risk will still be there. Disks may store, and CPUs may process, longer streams of historical data, but they will never alone guarantee risk-free return. Only networking speed and co-location seem to provide the possibility of falsifying the Efficient Market Hypothesis. On the other side, the obvious objection would be that such plusses could be replicated by anyone willing, and able, to invest enough time, resources and dedication to reach similar, or even better, results. In the end, the history of Mankind has always been one of today's losses and tomorrow's wins – the delta being the endeavour employed. Should not the main effort by regulators be the one to incentivise more HFT rather than less? To make competition easier for the many rather than harder for all? As suggested by Foresight (2012), "the more competitive the HFT industry, the more efficient will be the market in which they work" (ibid. p.54).

The question is still a very open one.

6.4. CONCLUSION ON HFT ARBITRAGE

The considerations developed above lead to the conclusion that, although many academics support the beneficial role of HFT in improving market efficiency, price discovery, and reduced transaction costs, the contrasting views on this matter are not devoid of good reasons. Because of their speed advantage, a relatively small number of HF traders are able to beat, most of the times, the majority of slower competitors in adjusting their own limit orders, cancelling them out and readjusting their price, or aggressively picking off other investors' outstanding limit orders. Moreover, using a simulation of arbitrage opportunities, this chapter demonstrated that, in contrast to the Efficient Market Hypothesis, a relatively small number of HF traders could consistently achieve risk-free return by reaping most of the arbitrage-led gains rather than letting them being spread evenly among the larger number of investors. Certainly, HF traders are speeding up price discovery - but to their own benefit and at the detriment of the many other slower investors, that is, at the detriment of the market at large. Yet, whether or not this violation of the market efficiency principles should be further regulated is still a debated matter. HF traders also have the chance to exploit the anomaly permitted by many exchanges' regulations of behaving like market makers when it suits them, profiting from the spread and taking advantage of liquidity providing rebates as incentives, without bearing any of the obligations market makers carry. As a last point, the debate on whether or not HF traders are the major contributors to decreasing transaction costs seems still rather open, and if many academics support the view of HFT decreasing transaction costs, there are other, not less authoritative, voices carrying a different opinion also worth listening to.

7. FLASH CRASH DATA ANALYSIS

7.1. INTRODUCTION

This chapter bridges the gap between the simulations and the audit trail data. The purpose is to provide hard data confirmation of some of the findings achieved through the theoretical approach. In particular, the data presented in the following sections are taken from the first week of May 2010, the one in which the Flash Crash occurred. Data from May 6, the critical day, are compared with data from the other days of the same week. Paragraph 7.2 shows a mathematical model of an exchange based on the Petri Nets, with the purpose to provide theoretical evidence of the role of Stop Loss orders in a declining market. Then, paragraph (7.3) analyses audit trail data from the Flash Crash to find confirmation about the exacerbating role of Stop Loss orders found in theory. Since no direct evidence can be inferred from the data, an indirect approach has been used. The hypothesis of a role of Stop Loss orders in exacerbating the volatility on May 6 is tested making use of the audit trail data. An original path of research concerns the so-called naïve orders (section 7.3.13), that is, those limit orders quoted at a price well above the top of the book, which have no rational justification unless those orders are assumed to be affected by a relatively large delay. The purpose of this section is to provide indirect confirmation of the findings about HFT-induced volatility, as resulted from the simulation discussed in chapter 4. Another testable hypothesis checked in paragraph 7.4 against real world data is whether frequent order cancellation may also contribute to exacerbating an already undergoing crisis. At that stage the concept of 'absolute speed' is introduced to explain market dynamics usually disregarded by the literature. The same mathematical model (Petri Nets) used at the beginning of the chapter, has been used again after discussing the role of order cancellations during the Flash Crash, to explain a mechanism compatible with the evidence obtained from the data analysis.

7.2. A SIMULATION USING PETRI NETS

7.2.1. A PETRI NETS MODEL OF AN EXCHANGE

As seen above, market participants behave in different ways, and so do market components. Order entry is typically random, checks and controls occur immediately as the appropriate conditions appear and some hardware or software sub-systems have intrinsic latencies. In the finite state machine representing a financial market some states are binary (yes/no, 0/1, true/false) whereas others have a more articulate blend of nuances. In order to formally describe a market using Petri Nets (PN) the basic definition of the latter does not fit: more advanced features are required. In particular a model representing a financial market, beyond places and transitions as in standard Petri Nets, requires: (i) input transitions, triggered randomly; (ii) latency-prone transitions; (iii) finite number of states, representing the price, associated with tokens that represent orders.

(i) Input transitions are random. An investor may enter or cancel orders at will, with no constraints at all regarding the time to do so.

(ii) Some non-input transitions are more latency-prone than others. The latency may be caused by external factors, like heavy traffic in a network, or internal ones, as behind-the-scene operations not explicitly represented in the Petri Nets but existing nevertheless. A practical example in the PN representing the Chicago Mercantile Exchange is the algorithm for working out whether a Stop Logic mechanism is to be triggered. That involves a prediction of the difference between the current price and the final one, should the current transaction, and all the stop loss orders triggered by it, be executed. Working out the scenario before allowing it to materialise requires time. Although the CME continuously strives to reduce this latency, it is undeniable that the latency will always be higher than in case no such check being carried out. Since no all modules in the PN are subject to similar behind-the-scene activities, they will all have different latencies. This is a feature the PN model has to take into account.

(iii) Each limit order has a price attached to it. It means that each token (representing an order flowing through the Petri Net) must have an attribute that specifies its price.

The Petri Net representing an exchange is depicted in figure 10.

At the beginning the system awaits for an input from the external world (the investors) by entering a bid (transaction 'bid') or an ask order (transaction 'ask') at a certain price. Such orders may come with stop loss ('bid SL' or 'ask SL') or without. In all cases the order can be cancelled until it stays on the order book ('cancel bid', 'cancel bid SL', 'cancel ask', 'cancel ask SL'). When both a bid and an ask order are present in the respective books, the transaction 'match bid-ask' can fire to check the possibility of a trade execution. If the two prices match then the transaction proceeds with the transition 'execute' and if neither order had a Stop Loss associated, the process is ready to start over again after firing the two ancillary transitions 'bin bid exec' and bin ask exec'. Had either or both limit orders a Stop Loss associated, the Stop Loss order gets activated (transition 'activate bid SL' or 'activate ask SL'). If at least one Stop Loss order is activated, first the Stop Loss price has to be checked against the last traded price, via the transition 'match bid SL' or 'match ask SL'. If the two prices do not match (that is, if the trading price is greater than the bid Stop Loss price, or less than the ask Stop Loss price) the 'bin bid-SL' or 'bin ask-SL' transition is fired and the process starts over again. Otherwise the transition 'book ask-SL' or 'book bid-SL' fires, transforming, respectively, the outstanding bid Stop Loss or ask Stop Loss order into an order in the appropriate book.

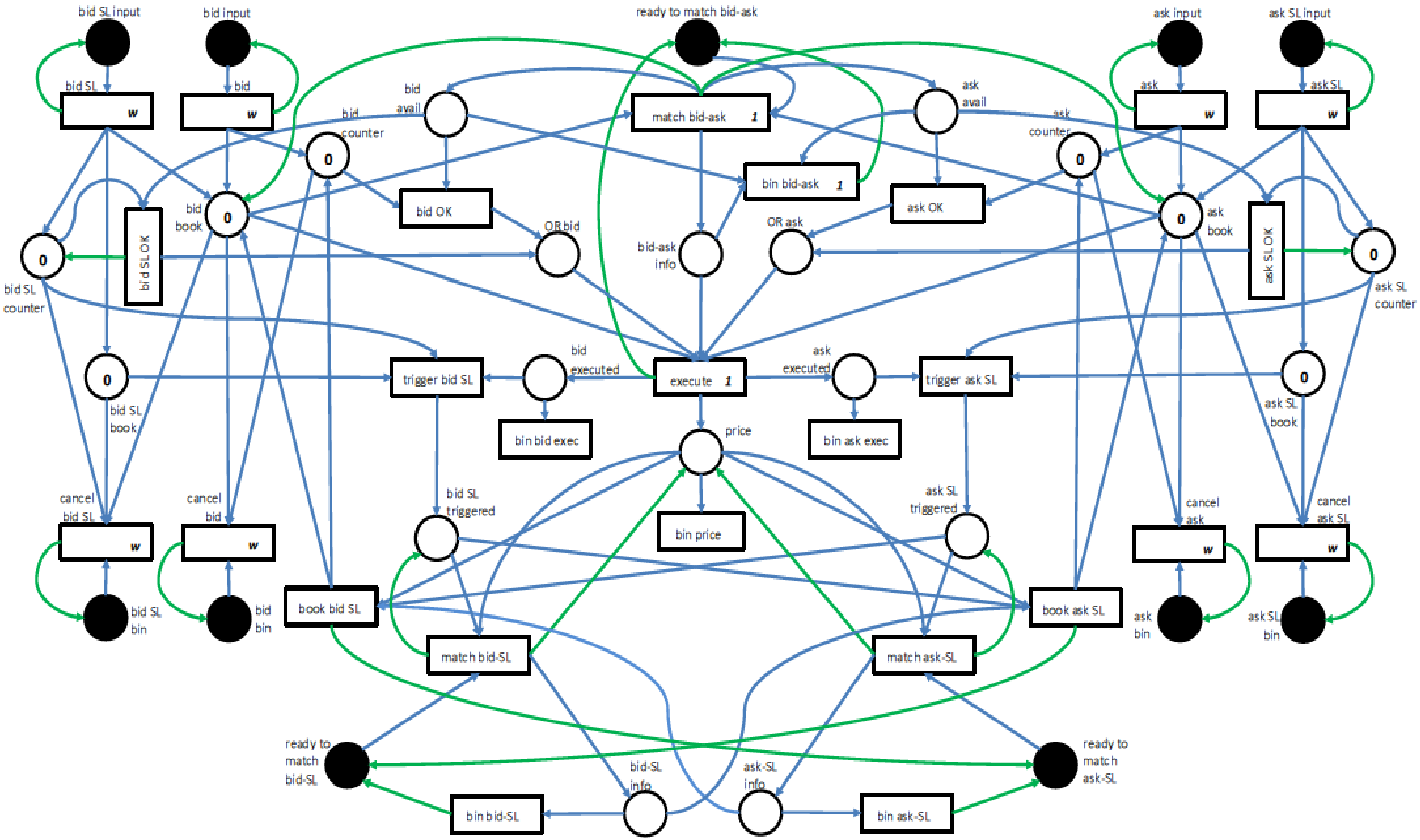


Figure 10. Petri Nets representing an exchange

7.2.2. MODELLING STOP LOSS ORDERS USING PETRI NETS

The dynamic behaviour of the exchange represented by the PN in figure 10 can be formally described by the step-by-step changes in the matrix associated with the PN. For purely illustrative purposes let us suppose the following example. Initially both the bid and ask books are empty. A 'bid SL' transition fires with bid price at, say, 87 and Stop Loss at 85. Then another 'bid' transition fires at 86. Other investors are more cautious and post bid orders by firing transitions 'bid' at 85 and 84, with no Stop Loss for sake of simplicity. On the other side of the book, an 'ask' transition fires at 87. Now transition 'match bid-ask' can fire and since the bid and ask price match, both the transitions 'bid SL OK' (since the bid order at 87 had a Stop Loss associated with it) and 'ask OK' fire, setting all the conditions for a transition 'execute' to fire. On the ask side there is no Stop Loss and so the transition 'bin ask exec' fires and nothing more happens on right-hand side of the PN. On the left-hand side the situation is more articulate. Since the bid order had a Stop Loss associated with it, the transition trigger bid SL' fires preparing the ground for the transition 'match bid-SL'. Since the trading price (87) and the Stop Loss price (85) are different, they do not match, transition 'bin bid-SL' fires and nothing more happens for the time being except the firing of the transition 'bin price', the only one available at this stage. At this point an ask order is posted at 86, which leads to transition 'match bid-ask' to fire against the highest bid quote, at 86. The two prices match, causing both 'bid OK' and 'ask OK' to fire, and leading to transition 'execute' to fire on its turn. This time neither the bid nor the ask side has Stop Loss, so the transitions 'bin bid exec' and 'bin ask exec' fire. Once again, the transition 'match bid SL' compares the trading price (86) and the Stop Loss price on hold (85) and since they do not match, 'bin bid SL' and 'bin price' fire, cleaning up the situation. Eventually an ask order at 85 arrives (transition 'ask'). The presence of orders on both the bid and ask books triggers the transition 'match bid-ask', which compares the highest bid price with the lowest ask price. They are both at 85 so there is a match (transitions 'bid OK' and 'ask OK') which leads to an 'execute'. None of the orders just executed had a Stop Loss associated, so transitions 'bin bid exec' and 'bin ask exec' fire. On the right-hand side the transition 'match bid SL' finds that the trading price at 85 is equal to the Stop Loss price, causing 'book ask SL' to fire, which transforms the Stop Loss order into an ask market order. Now there is a limit order at 84 on the bid side and a market order on the ask side. Although the latter was originally posted at 85, given that there is no limit order at 85, liquidity at that price level has been consumed, when transition 'match bid-ask' fires, the two orders match, followed by a 'bid OK' and an 'ask OK'. Therefore transition 'execute' fires, to the disappointment of the first investor that wished to partially protect its long position, opened at 87, with an exit strategy at 85 and instead has to suffer a further loss by selling at 84. The example is deliberately quite simple as in a real exchange there are many more outstanding limit and Stop Loss, orders at any one time. However, for the purpose of studying the market behaviour, the scarce liquidity at 85 caused a Stop Loss order to exacerbate volatility in an already falling market. Potentially, triggering of Stop Loss orders may consume the available liquidity and in case of decreasing liquidity, leading to price jumps and negative feedback loops.

The scenario presented above is purely theoretical but the premises are all there, namely a nose-diving market, scarce and decreasing liquidity - the consequence being higher-than expected volatility. This suggests that something similar might have occurred on the day and time of the Flash Crash, or at other times. But before turning this suggestion into a more concrete statement, it must be checked whether audit trail data supports this conclusion. This is the purpose of the next paragraph.

7.3. ROLE OF STOP LOSS ON THE FLASH CRASH

7.3.1. INTRODUCTION

As seen in the previous paragraph stop loss orders have, at least in theory, the potential to trigger negative feedback loops and this may also have had a role in the Flash Crash. In order to verify this hypothesis, detailed data would be necessary, showing which limit orders had a Stop Loss order associated with them and at which price. Unfortunately, this level of details is not publicly available. The Chicago Mercantile Exchange sells market data messages needed to recreate the 10-level top of book for products traded on the CME Globex electronic trading platform. That includes all changes to the book including bids, asks, bid volumes and ask volumes - ten levels deep time-stamped to the millisecond. However, since a detailed analysis about the impact of Stop Loss orders is not possible in a direct way, a different approach is required. The following section shows the data and illustrates the methodology used throughout this paragraph. The following sections show the results suggested by the data. The main hypotheses will then be tested by using different statistical techniques and, before discussing the overall findings (section 7.3.14), the original concept of naïve orders shall be presented (in section 7.3.13). Section 7.3.15 concludes the paragraph.

7.3.2. METHODOLOGY

In order to evaluate the impact of Stop Loss orders onto a crisis, observation cannot be done directly as, commercially available data does not contain information about whether a market order is originated by a Stop Loss order or otherwise. Therefore the only way to work it out is to use a proxy and the proxy that has been used in the following sections is the length of a 'run'. A 'run' is defined as an uninterrupted sequence of trades all in the same direction (that is, aggressive orders all against a bid quote or all against an ask quote); the 'length' of the run is the number of trades within a run. The Flash Crash model developed by Aldridge (2014) identifies trade runs and thus it indirectly detects aggressive order patterns. The length of the run is driven by aggressive order flow and the author finds a correlation between one-sided flow and future market volatility. Obviously there is no absolute guarantee that trades in a run are caused by a sequence of Stop Loss orders. An example is given by the run which occurred on the day before the Flash Crash. A run of length 8 started at 18:45:51 and 319 milliseconds Greenwich Mean Time (GMT, that means 2:45:51.319pm in New York or one hour earlier in Chicago) on the ask book of the E-mini S&P 500 futures contracts traded at the Chicago Mercantile Exchange. At each step of the run only one contract was traded. The second trade occurred at the millisecond 332, then one at 337, at 352, 358, 375, 378 and 394. The total time covered by the run was 75 milliseconds and this can hardly configure a sequence of Stop Loss orders as they are normally launched with no latency in-between. In this case the latencies between successive trades were 13 milliseconds, then 5ms, then 15, 6, 17, 3, and 6. In another case on the same day and on the same book, a run of length 8 executed within one millisecond; the run started and terminated at 18:27:27.115, initially trading 2 contracts, then another 2, then 2, 2, 2, 3, 2 and 1. Seventy-five milliseconds versus one: it sounds sensible to identify quite some HFT activity (including time spent in decision-making) in the former case and automatic Stop Loss execution in the latter. In runs much longer than average, especially if executed in a very short time, it can be sensibly assumed that a special automatic mechanism was in place, and the typical automatic mechanism that can increase the length of a run in a short time is the execution of Stop Loss orders. It must be noticed that long runs do not necessarily cause large price movements or trouble. The second longest lasting run on May 5, 2010 (which lasted 72 milliseconds) displays a length of 684, trading a total of 1,248 contracts. The total price rise (as it was also on the ask book) was 0.25 index points, equal to one tick. The investigation has been carried out with the purpose to identify abnormal values in periods of time showing a comparable number of market events over different days and, once such anomalies have been identified, to understand the underlying causes. The date and time under primary observation are the six minutes and twenty-eight seconds (18:39:00.007 GMT through 18:45:28.115 GMT on 6/5/2010) leading to the triggering of the Stop Logic by the CME Globex platform, an event that started the recovery on the E-mini S&P 500 futures contract market. In order to evaluate such observations, those data will be compared with the same amount of data including the same six and a half minutes in the three previous days and in the following one; it means that the data taken on May 6, 2020 will be compared with the data from May 3rd, 4th, 5th and 7th. The criterion chosen for deciding the length of the investigation period was the number of records produced by the CME Globex platform in that period. This way it is sure that the same number of market events will be taken into account, for all the days observed. Since May 6 was a rather busy day, the same number of market events occurred during those six and a half minutes on that day needed a much longer period of time to occur on the other days. The findings of this investigation are reported in table 12. The header shows, for each day being investigated, the observation's time, its duration, the number of events. Then, Panel A displays the number of trade runs, the maximum, the average length and standard deviation of run lengths, and the number of runs whose length was greater than 300, 200, 100, 50, 25, and 10 trades, respectively. Panel B details, for each individual run, the difference between the initial price and the price at the end of the run. It then shows the maximum price difference, the average and the standard deviation, the number of runs in which the price difference was greater or equal than 3.25 index points (equal to 13 ticks), 3 (12 ticks), 2 (8 ticks), one index point (4 ticks), greater or equal than 0.5 index points (two ticks), and how many times at least one tick change occurred during a run, i.e. greater than zero. Panel C takes into account the number of order cancellations at the top of the book occurring within 10 milliseconds, within 3ms and within one millisecond after a run.

The following sections discuss the results highlighted by the analysis in particular focusing on those that mark May 6, 2010 as a very special day, with the goal of identifying the reasons that led to such an extreme outcome.

7.3.3. TRAFFIC

The first thing to notice is that, in order to analyse an identical number of market events, very different

Date	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
Time window	173318.954- 192748.770	181107.807- 191047.481	183348.932- 192318.863	183900.007- 184528.115	182906.396- 190434.858
Duration	01:54:30	00:59:40	00:49:30	00:06:28	00:35:28
Events	580,684	580,684	580,684	580,684	580,685
PANEL A					
Trade runs	11399	9293	8843	12824	6656
Max length	273	260	239	324	156
Avg length	2.91876	3.40170	3.37114	2.79141	3.10592
Stdev length	9.64347	10.72825	11.47181	8.71112	7.94710
Length > 300	0	0	0	1	0
Length > 200	3	2	4	4	0
Length > 100	25	25	26	19	6
Length > 50	87	90	103	63	39
Length > 25	219	203	219	184	137
Length > 10	503	543	438	575	380
PANEL B					
Delta price					
Max	0.25	0.25	0.25	3.25	0.50
Avg	0.00013	0.00073	0.00074	0.01482	0.00240
Stdev	0.005734387	0.013456604	0.013536676	0.110186533	0.025156276
>= 3.25 ip	0	0	0	1	0
>= 3 ip	0	0	0	4	0
>= 2 ip	0	0	0	10	0
>= 1 ip	0	0	0	27	0
>= 0.5 ip	0	0	0	109	2
> 0 ip	6	27	26	476	62
PANEL C					
Cancellations					

within 10ms	11690	9537	9885	15542	11106
within 3ms	4830	3994	4533	8013	4342
within 1ms	2359	2100	2176	4125	2059

Table 12. Data about trade runs (bid book)

periods in time had to be selected. Namely, the same number of events (580,864) which took place on May 6 during the six minutes, 28 seconds and 108 milliseconds leading to the halt, needed one hour 54 minutes and a half on May 3, one hour less twenty seconds on May 4, forty-nine minutes and a half on the 5th, and 35 minutes and 28 seconds on the 7th. The last row in table 13 shows that in less than six-and-a-half minutes on May 6 there were as many as six to nineteen times the traffic experienced on the other days (computed as no. of events on May 6 / no. of events on that day).

Date	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
Start time	18:39:00.006	18:38:59.819	18:39:00.003	18:39:00.007	18:39:00.005
Start event	2669181	9128697	16420889	25564917	42721111
End time	18:45:28.129	18:45:28.336	18:45:28.117	18:45:28.115	18:45:28.215
End event	2699963	9190806	16488885	26145764	42819161
# of events	30,782	62,109	67,996	580,684	98,050
Comparison	18.87	9.35	8.54	1.00	5.92

Table 13. Number of events

7.3.4. NUMBER OF TRADE RUNS

Also the number of trade runs which occurred on the day of the Flash Crash was greater than the number of runs on the other days, the increase ranging from 12.5% (compared to May 3) to nearly 93% (compared to May 7), as shown in table 14.

Date	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
Runs	11399	9293	8843	12824	6656
Runs increase on May 6 wrt	12.50%	38.00%	45.02%		92.67%

Table 14. Comparison between number of runs

This is suggestive of a higher-than average number of Stop Loss orders being executed on that day. It can be argued that this is an expected occurrence on a very volatile day, since wide price movements tend to trigger the Stop Loss mechanism more often. Nevertheless, the wide range for this indicator suggests that the Stop Loss mechanism is definitely a potential candidate to be one, albeit perhaps not the only one, of the main factors that contributed to the crisis. Moreover, the frequency of such events is also worth discussing (table 15).

Date	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
Runs	11,399	9,293	8,843	12,824	6,656
Duration	01:54:30	00:59:40	00:49:30	00:06:28	00:35:28
Runs/sec	1.65924	2.59581	2.97744	33.05155	3.12782
Sec/run	0.60268	0.38524	0.33586	0.03026	0.31971

Table 15. Run rates

On May 3rd, 4th, 5th, and 7th the run-per-second rate ranges between 1.7 (one event every 603ms, on the 3rd) and 3.1 (one event every 320ms, on the 7th), whereas the rate observed on May 6, that is 12,824 runs in 6 minutes and 28 seconds, is equal, on average, to 33 trade runs per second, or a run every 30 milliseconds. If in the non-Flash-Crash days the run rate, although very high, is at some extent still understandable by a trained human brain, the frequency on the 6th is far too high for even a human eye to grasp.

7.3.5. MAXIMUM AND AVERAGE LENGTH OF A RUN

On May 6, the maximum length of a run was highest (over the five days) at 324, versus a lowest maximum length of 156 on May 7 and an average of 250 for the days other than the 6th. Under the reasonable assumptions that a large number of trades occurring within a few milliseconds (and in many cases the whole run occurred within the same millisecond) cannot be but automatic, it looks very likely that after the initial (few) investor-driven trades within a run, a rather large number of Stop Loss orders were executed in sequence. Although the number of trade runs and their respective maximum lengths are clearly different, their means are close to each other (as shown in table 12 – panel A). Therefore, it makes sense to investigate whether the means are significantly different from each other. The technique used throughout this investigation is the analysis of variances, or ANOVA for short. Table 16 summarises the steps required by the ANOVA technique for comparing the means of the trade runs' lengths, where the naming convention in the first column follows the one adopted by Trivedi (2016).

	Length of trade runs				
	03-May	**04-May**	**05-May**	**06-May**	**07-May**
Mean of group	2.91876	3.40170	3.37114	2.79141	3.10592
Error (within groups)	1059975	1069465	1163628	973055	420306
Group size (n)	11399	9293	8843	12824	6566
Total mean (SST)	3.08404				
Between treatments	311.35688	937.76467	728.92475	1098.13954	3.18761
SS(Tr) (sum of squares between treatments)	3.079.37				
SSE (error sum of squares)	4686429.48				
Degrees of freedom between treatments (c-1)	4				
Degrees of freedom error (NT-c)	49010				
SS(Tr)/(c-1)	769.84				
SSE/(NT-c)	95.62				
F-statistics	**8.05**				

Table 16. ANOVA for the length of trade runs

The result of the F-statistic (8.05) is clearly larger than the value of $F_{(4,49015)}$ = 2.37, leading to conclusion that at least two means of trade runs are different.

7.3.6. NUMBER OF RUNS LONGER THAN 10 TRADES

It is interesting to notice that the average length of the runs shown in table 12 was lowest on May 6 (2.79 versus 2.92, 3.40, 3.37 and 3.10, respectively on the 3rd, 4th, 5th, and 7th), suggesting that most runs on that day were quite short while the peaks were very long. Consistent with this remark, it can be observed that table 12 shows the number of runs respectively longer than 100, 50, and 25 trades was not highest on May 6, which instead showed a high in the number of runs longer than 300 trades and longer than 10 trades. This is relevant under the quite reasonable assumption that Stop Loss order triggering definitely occurred at least in runs longer than 10 trades. This observation suggests that on the Flash Crash day the runs had extreme behaviours: a small number of very long runs and a larger than average number of short runs, showing a below-average number of medium-length runs. This observation again confirms the erratic behaviour of the market, at least during those critical six-and-a-half minutes.

7.3.7. MAXIMUM PRICE DROP WITHIN A RUN

By far the most interesting finding is the maximum price difference within a run: on May 6 it was 3.25 index points (equivalent to USD 162.50) versus 0.25 index points (USD 12.50) on the 3rd, 4th and 5th, and 0.5 index points (USD 25.00) on the 7th. This provides clear evidence: Stop Loss orders were likely contributors to the dramatic price drop on the day and at the time of the Flash Crash but it could not have happened if the long runs triggered by Stop Loss orders had not hit the vacuum, forcing the price to match sequentially downward to the next price level. At 18:45:13.871 GMT a run of length 99 started and went on for 10 milliseconds, until 18:45:13.881. It caused the price of the E-mini S&P 500 futures contracts June 2010 falling down thirteen ticks (from 1074.50 down to 1071.25), while price change was usually restricted to one or two ticks at the maximum on the other days. The dollar amount involved in that run exceeded twenty-one millions. All large price movements occurring on May 6 during a run are summarised in the table 17.

The run encompassing the second largest price movement, 3 index points or 12 ticks, in two cases lasted 2 milliseconds and in one case lasted 7 milliseconds. Only two runs which experienced large price movements, of 2.25 index points, lasted more than 10ms (the one started at 18:45:12.444

lasted 15ms and the one commenced at 18:45:17.954 lasted 16ms), whereas in two cases the run started and terminated within the same millisecond. In particular, the run starting at 18:45:12.960 GMT had a length 2 and a price change of 2.5 index points (10 ticks).

START TIME	INIT PRICE	END TIME	LAST PRICE	TRADES	DELTA PRICE	MILLI-SECONDS
184513871	1074.5	184513881	1071.25	99	3.25	10
184512489	1074.75	184512491	1071.25	21	3.00	2
184526518	1070.25	184526525	1067.25	83	3.00	7
184527996	1066	184527998	1063	16	3.00	2
184512952	1075	184512960	1072.25	82	2.75	8
184518693	1071.75	184518699	1069	68	2.75	6
184512960	1075	184512960	1072.5	2	2.50	0
184512444	1075	184512459	1072.75	125	2.25	15
184517945	1075	184517961	1072.75	126	2.25	16
184528111	1059	184528114	1056.75	31	2.25	3
184511702	1076.75	184511709	1075	65	1.75	7
184526894	1064.75	184526897	1063	30	1.75	3
184528107	1062	184528109	1060.25	28	1.75	2
184508584	1077.5	184508589	1076	58	1.50	5
184511744	1076	184511748	1074.5	39	1.50	4
184522305	1068	184522307	1066.5	19	1.50	2
184527018	1064.5	184527018	1063	7	1.50	0
184458528	1081.75	184458532	1080.5	46	1.25	4
184518699	1069	184518705	1067.75	29	1.25	6
184518705	1067.75	184518707	1066.5	24	1.25	2
184521096	1070.75	184521098	1069.5	22	1.25	2
184521415	1068	184521417	1066.75	30	1.25	2
184521629	1068	184521630	1066.75	9	1.25	1

Table 17. Largest price movements on May 6, 2010 (ordered by descending Delta Price)

Indeed, what happened in that case confirms that the scenario described in chapter ‘4. Impact of High-Frequency Trading on Volatility’, was founded on a solid basis. At 18:45:12.952 GMT the prevailing bid price was 1075 index points. At that time a run started and in just 8 milliseconds it consumed all the liquidity available down to 1072.25. Just after that, at 18:45:12.960, a bid was quoted at 1075, which got gobbled immediately, causing a large quote drop on the bid book in no time. A sensible explanation is that an investor, probably a computer that had a small but not negligible latency, noticed the prevailing bid at 1075 at, or before, 18:45:12.952, launched its own bid at that price but, because of its latency, the quote only arrived at 18:45:12.960, after all the liquidity at 1075 (and down to 1072.25) had been taken away. In the words of the simulation described in chapter 4, its order was 'queued' for at least 8ms (for example because the computer was not co-located). The result was that when the bid order at 1075 hit the book, it found itself totally isolated from other bid limit orders which were, at that time, quoting at most 1072.25. The lonely quote was immediately taken by a lucky HF trader which sold 2.75 index points above the prevailing bid price just exploiting the other trader's latency and its own rapidity. So, not only latency-prone market orders are at risk of nasty surprises but limit orders too could (and did) become stale in a matter of a few milliseconds - even before they reach the exchange server. In other words, in the Age of HFT, orders may become obsolete between their conception and the time they are born. This phenomenon will be investigated in more depth in section ‘7.3.13. Naïve Orders’.

A variance analysis (ANOVA) has also been performed for the price change (table 18).

	Delta price within a run				
	03-May	**04-May**	**05-May**	**06-May**	**07-May**
Mean of group	0.00013	0.00073	0.00074	0.01482	0.00240
Error (within groups)	0.37480	1.68260	1.62022	155.68497	4.21154
Group size (n)	11399	9293	8843	12824	6566
Total mean (SST)	0.00450				
Between treatments	0.21790	0.13260	0.12560	1.36374	0.02935
SS(Tr) (sum of squares between treatments)	1.87				
SSE (error sum of squares)	163.57				
Degrees of freedom between treatments (c-1)	4				

Degrees of freedom error (N_T-c)	49010
SS(Tr)/(c-1)	0.47
SSE/(N_T-c)	0.00
F-statistics	**140.01**

Table 18. ANOVA for the price change

Even in this case the F-statistics yields a result which clearly indicates a mean difference between the five populations. In both the length of trade runs (table 16) and in this case, the value corresponding to May 6th is at one extreme of the range of values. But whereas in the analysis of variance of trade runs' length the F-Statistic is somehow larger than three times the value of $F_{(4,49015)}$ (that is, 8.05 versus 2.37), in the delta price variance analysis the ratio is nearly sixty times larger (140.01 versus 2.37), as the means themselves suggest (0,01482 on May 6 versus 0,000013 on May 3). The price change within a run on the day of the Flash Crash was definitely an abnormal event.

7.3.8. AVERAGE DELTA PRICE OVER ALL RUNS

The price difference between the beginning and the end of a run is indicative of the dramatic price drop, very likely caused by the Stop Loss mechanism. But the peak (3.25 index points on May 6 against 0.25 or 0.5 on the other days) tells only half of the story. Table 19 shows the average of the price difference across the runs for each of the days under observation. The following three rows show the ratio between the averages normalised on May 3, May 4, and May 7, respectively. It means that, for example, on the fourth row the average for May 3 has been taken equal to 1 and the others as a multiple of that average. So, on May 4 and May 5 the average was about 5.5 times higher than on the 3rd, on May 7 it was 18 times higher and on the 6th the ratio was more than 112. The fifth row (only ratios greater than one are displayed) shows that on May 6 the average was more than 20 times higher than on the 4th and more than 6 times than on the 7th (sixth row).

	DELTA PRICE				
Date	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
Avg	0.00013	0.00073	0.00074	0.01482	0.00240
	1	5.51980	5.58585	112.59150	18.26763
		1	1.01197	20.39775	3.30947
				6.16344	1

Table 19. Average price difference normalised

The actual figures are not so important (as they vary considerably) as the qualitative indication: on the day of the Flash Crash the average price drop during a run was very much higher than on any other day.

7.3.9. FUTHER HYPOTHESIS TESTING

The ANOVA technique only allows to spot whether at least two means, among all the data sets, are different, without indicating how many and which ones. Therefore, in order to surpass both uncertainties, hypothesis testing has been performed on every possible pair of days under investigation. Table 20 shows, for both the length of trade runs (in panel A) and for the price change per each run (in panel B), the Z-statistics on the day-over-day comparison.

PANEL A	**Z-values**	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
	03-May		-3.36940***	-2.98026***	1.07348	-1.40885
	04-May	3.36940***		0.18506	4.51111***	1.99990**
	05-May	2.98026***	-0.18506		4.01979***	1.69893*
	06-May	-1.07348	-4.51111***	-4.01979***		-2.53393**
	07-May	1.40885	-1.99990**	-1.69893*	2.53393***	
Length of trade runs	Avg	2.91876	3.40170	3.37114	2.79141	3.10592
	StDev	9.64347	10.72825	11.47181	8.71112	7.94710
	Size	11399	9293	8843	12824	6656
PANEL B	**Z-values**	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
	03-May		-3.97655***	-3.92762***	-15.06878***	-7.25984***

	04-May	3.97655***		-0.04335	-14.33371***	-4.95607***
	05-May	3.92762***	0.04335		-14.31571***	-4.90401***
	06-May	15.06878***	14.33371***	14.31571***		12.16043***
	07-May	7.25984***	4.95607***	4.90401***	-12.16043***	
Delta price per run	Avg	0.00013	0.00073	0.00074	0.01482	0.00240
	StDev	0.00573	0.01346	0.01354	0.11019	0.02516
	Size	11399	9293	8843	12824	6656

Table 20. Z-statistics comparing day over day

The values marked with one asterisk (*) indicate the day pairs for which the null hypothesis of equal means can be rejected at 90% confidence level but cannot be at 95% confidence level; those marked with two asterisks (**) indicate day pairs for which the means equality can be rejected at 95% but not at 99%, and the three asterisks (***) mean day pairs for which means equality can be rejected at 99% or higher confidence level. For values with no asterisks the null hypothesis can be rejected at 90% or lower confidence level.

Both tables are anti-symmetrical with respect to the main diagonal top left to bottom right (since the Z-value resulting from the mean difference between, say, May 3 and May 4 has the same absolute value, and the opposite sign, as the mean difference between May 4 and May 3). Prerequisites for parametric tests, like Z-statistics or the analysis of variance, to yield acceptable results are the specific assumptions that can be made about the distribution of the population(s) from which the data sets are extracted. In chapter 6, the Bowman-Shelton's normality of distribution test has been used for checking the validity of the Z-statistics assumptions. Other cases are more complex. In particular, one of the prerequisites for using ANOVA is the same variance for all the populations from which the data sets are extracted. This is a difficult condition to enforce. Population variances are often not known, and even if they were, being them all equal would be a very unlikely occurrence. In such cases, statistics rules dictate the use of non-parametric tests. In general, there is a consistent agreement among statisticians to use parametric tests, unless clear evidence exists that the underlying assumptions have been violated. The opposite case is more debated: not such large consensus exists about the need to use non-parametric tests in case of violation of the assumptions [Sheskin (2004)]. Indeed, according to some sources, the higher power carried by parametric tests is a strong reason to adopt them with some degree of liberality, namely even when one or more assumptions are violated at a reasonably slight degree. Other statisticians agree that the lesser power of non-parametric tests is compensated by their wider validity, although these tests tend to sacrifice information at some extent. The 'liberal' approach relies on the robustness of most parametric tests, where a 'robust' test is one that still provides reasonably reliable results even though not all of its assumptions hold. The conclusion reached by Sheskin (2004) is that "in most instances, the debate concerning whether one should employ a parametric or nonparametric test to evaluate data derived for a specific experimental design turns out to be of little consequence. The reason for this is that most of the time a parametric test and its nonparametric analog are employed to evaluate the same set of data, they lead to identical or similar conclusions" (ibid. p.127). That said, there remain to check whether in this case parametric (ANOVA) and non-parametric (Kruskal-Wallis) tests yield 'identical or similar' conclusions. If so, the degree of confidence in the reliability of results can be ascertained to be high enough.

7.3.10. THE KRUSKAL-WALLIS TEST

The null hypotheses tested with the Kruskal-Wallis technique are the same tested with ANOVA earlier: the means of (i) trade run lengths and (ii) price change for the five populations (shown in table 12), whose samples have been collected on May 3 through May 7, 2010 are all equal, against the alternative hypotheses that at least one of the means for the trade run lengths, and at least one for the price changes is different from the others. The notation used throughout this section is taken from Sheskin (2004).

$H_0 : \theta_{3\text{-May}} = \theta_{4\text{-May}} = \theta_{5\text{-May}} = \theta_{6\text{-May}} = \theta_{7\text{-May}}$, and

H_1 : not H_0, i.e., at least one $\theta_{i\text{-May}}$ is different from the others (with 3 <= i <= 7).

The size of the samples are as reported in the rows Size' at the bottom of Panel A or Panel B in table 20, where the total size of the five samples is 49,015. The sums and averages of the ranks for the trade run lengths are as displayed in table 21.

RUN LENGTH			DELTA PRICE		
DAY	SUM	AVG	DAY	SUM	AVG
03-May	271181890.00	23,789.97	03-May	276110802.50	24,222.37
04-May	230085655.00	24,759.03	04-May	225639087.50	24,280.54
05-May	213339559.00	24,125.25	05-May	214720360.50	24,281.39
06-May	317557693.50	24,762.76	06-May	322134373.50	25,119.65
07-May	169094822.50	25,404.87	07-May	162654996.00	24,437.35

Table 21. Sum and average ranks for run lengths and price changes

Then applying the Kruskal-Wallis test to the run length analysis, H = 69.65

As usual in this test, the value of H can be compared to χ^2 in the Chi-squared test, in this case with 5 - 1 = 4 degrees of freedom. Since H is greater than $\chi^2_{.001}$ = 18.47, it can concluded that there is at least one mean different from the others.

The same approach can be used for testing whether the means of the price changes are all equal (null hypothesis) or at least one is different (alternative hypothesis). In this case H = 33.44

Again, comparing H to χ^2 with df=4, H > $\chi^2_{.001}$.

In both the analysis of the trade runs length and the price change cases, the outcome of the Kruskal-Wallis test is similar to the one obtained with the ANOVA technique. The most noticeable difference between the two techniques is that with ANOVA the result for the case of price change was more than 17 times larger than the result for the case of run lengths, whereas with the Kruskall-Wallis test the latter is less than half of the former. Since the ratios among the means for the trade runs length ranged within a 22% band and those for the price change went as up as 1,930%, it is confirmed that, although the qualitative results are 'similar', the non-parametric test does definitely sacrifice information with respect to the more powerful parametric test.

7.3.11. NUMBER OF RUNS SHOWING PRICE JUMPS

Another interesting suggestion comes from the number of runs showing a certain price drop. Whereas there is no point in comparing price drops higher than 0.5, as only May 6 shows a number greater than zero in those cases, it is interesting to compare the number of runs experiencing at least one tick price change across the days under scrutiny, as in table 22, where the second row displays the number of runs experiencing price changes in the periods under observation.

	DELTA PRICE				
Date	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
> 0 ip	6	27	26	476	62
	1	4.50000	4.33333	79.33333	10.33333
		1.03846	1	18.30769	2.38462
				7.67742	1

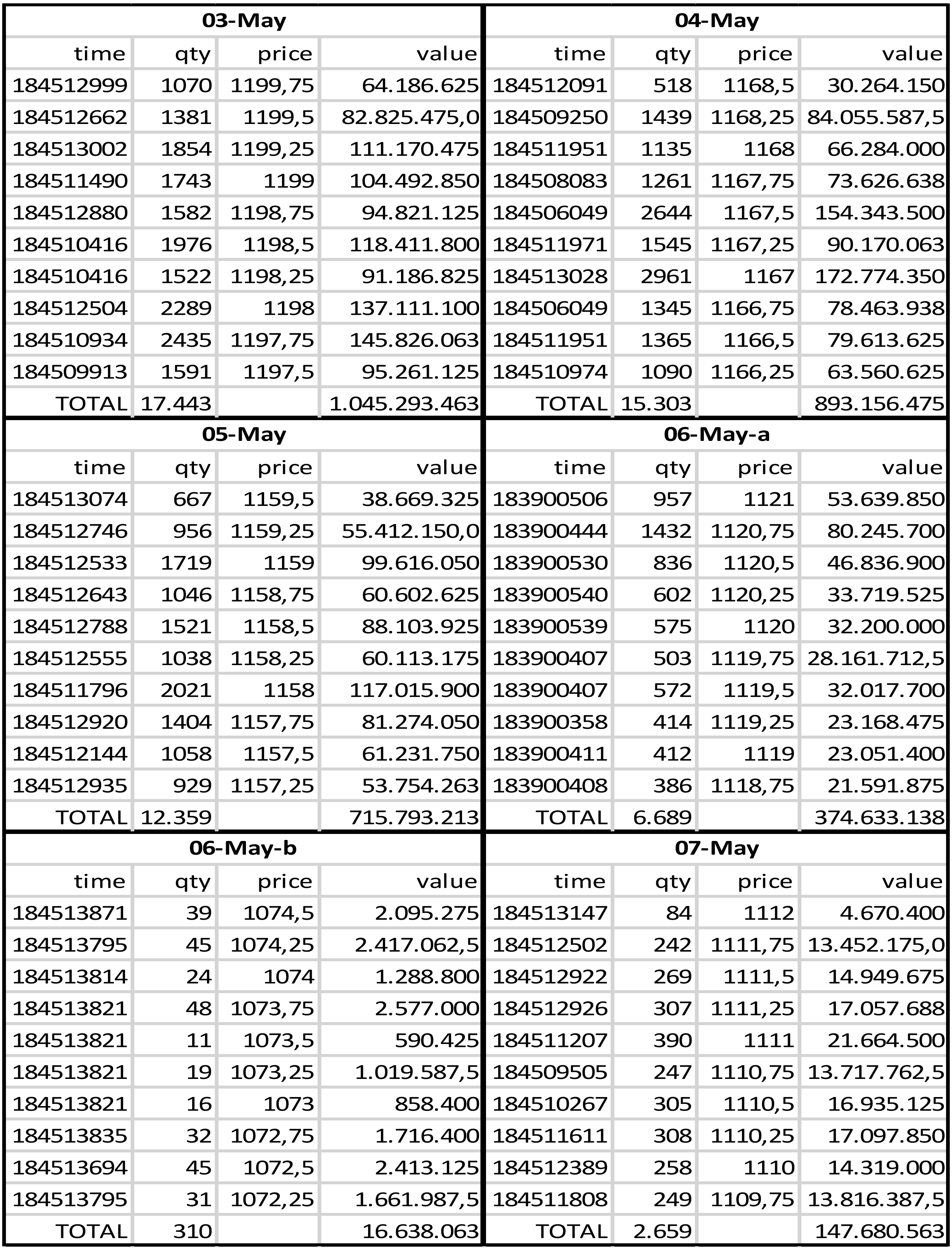
Table 22. Absolute and normalised price change

Taking May 3 as a reference, the third row shows that on the 4th and the 5th there were, respectively, four and a half and four and one third as many runs during which the price changed by at least one tick, more than 10 times on May 7 and more than 79 times on the 6th. The comparison between May 6 and the second most volatile day (May 7) shows a ratio as high as 7.677 times (last row). So, the price changes during a run clearly shows that on the Flash Crash day price drops caused by Stop

Loss orders were not only on average larger in size per each run, but they also occurred much more frequently.

7.3.12. LIQUIDITY

Liquidity has been considered one of the biggest issues of the Flash Crash. The combination of price uncertainty and withdrawal of many market makers and other liquidity suppliers led liquidity on that day to virtually vanishing. To better understand the extent to which this phenomenon materialised, table 23 compares liquidity available at the ten top levels of the bid book at the most critical time on May 6 (heading: 06-May-b), compared to the time when the crisis started on the same day (heading:



03-May				04-May			
time	qty	price	value	time	qty	price	value
184512999	1070	1199,75	64.186.625	184512091	518	1168,5	30.264.150
184512662	1381	1199,5	82.825.475,0	184509250	1439	1168,25	84.055.587,5
184513002	1854	1199,25	111.170.475	184511951	1135	1168	66.284.000
184511490	1743	1199	104.492.850	184508083	1261	1167,75	73.626.638
184512880	1582	1198,75	94.821.125	184506049	2644	1167,5	154.343.500
184510416	1976	1198,5	118.411.800	184511971	1545	1167,25	90.170.063
184510416	1522	1198,25	91.186.825	184513028	2961	1167	172.774.350
184512504	2289	1198	137.111.100	184506049	1345	1166,75	78.463.938
184510934	2435	1197,75	145.826.063	184511951	1365	1166,5	79.613.625
184509913	1591	1197,5	95.261.125	184510974	1090	1166,25	63.560.625
TOTAL	17.443		1.045.293.463	TOTAL	15.303		893.156.475

05-May				06-May-a			
time	qty	price	value	time	qty	price	value
184513074	667	1159,5	38.669.325	183900506	957	1121	53.639.850
184512746	956	1159,25	55.412.150,0	183900444	1432	1120,75	80.245.700
184512533	1719	1159	99.616.050	183900530	836	1120,5	46.836.900
184512643	1046	1158,75	60.602.625	183900540	602	1120,25	33.719.525
184512788	1521	1158,5	88.103.925	183900539	575	1120	32.200.000
184512555	1038	1158,25	60.113.175	183900407	503	1119,75	28.161.712,5
184511796	2021	1158	117.015.900	183900407	572	1119,5	32.017.700
184512920	1404	1157,75	81.274.050	183900358	414	1119,25	23.168.475
184512144	1058	1157,5	61.231.750	183900411	412	1119	23.051.400
184512935	929	1157,25	53.754.263	183900408	386	1118,75	21.591.875
TOTAL	12.359		715.793.213	TOTAL	6.689		374.633.138

06-May-b				07-May			
time	qty	price	value	time	qty	price	value
184513871	39	1074,5	2.095.275	184513147	84	1112	4.670.400
184513795	45	1074,25	2.417.062,5	184512502	242	1111,75	13.452.175,0
184513814	24	1074	1.288.800	184512922	269	1111,5	14.949.675
184513821	48	1073,75	2.577.000	184512926	307	1111,25	17.057.688
184513821	11	1073,5	590.425	184511207	390	1111	21.664.500
184513821	19	1073,25	1.019.587,5	184509505	247	1110,75	13.717.762,5
184513821	16	1073	858.400	184510267	305	1110,5	16.935.125
184513835	32	1072,75	1.716.400	184511611	308	1110,25	17.097.850
184513694	45	1072,5	2.413.125	184512389	258	1110	14.319.000
184513795	31	1072,25	1.661.987,5	184511808	249	1109,75	13.816.387,5
TOTAL	310		16.638.063	TOTAL	2.659		147.680.563

Table 23. Comparison of liquidity

06-May-a) and to about the same time on the other days. The aggregate top ten level liquidity varies a lot over the range investigated. However, it seems reasonable to take the liquidity shown on the 3rd, the 4th and the 5th as somehow standard, because May 6 was affected by very negative news since the beginning of the trading day and because 18:39 was already a critical time for the E-mini S&P 500 futures contracts. Moreover, May 7th can hardly be considered a standard day as the market was likely still rather shocked by the previous day's events and, understandably, liquidity suppliers were much more cautious than usual. As shown in table 24, liquidity on the most critical time on May 6 decreased 98% with respect to the same time on May 3 and May 4, more than 97% with respect to May 5, 95% compared to just six minutes earlier, and still an appalling 88% with respect to the following day. Overall, on May 6 the combination of Stop Loss orders and scarce liquidity had an enormous impact on the price drop: the runs were on average much longer than on the other days, there were more and higher peaks, and even the short runs were larger in number; the beginning of a run was often the alarming start of a dramatic downward price movement, such movements were large and frequent, and downticks at millisecond level was a common occurrence.

As far as the Flash Crash is concerned, it is worth noticing that all, except one, price movements larger than one index point occurred in the last 20 seconds before the halt (table 17).

Date	**03-May**	**04-May**	**05-May**	**06-May-a**	**06-May-b**	**07-May**
Liquidity	17443	15303	12359	6689	310	2659
Loss of liquidity wrt	98.22%	97.97%	97.49%	95.37%		88.34%

Table 24. Loss of liquidity on May 6

7.3.13. NAÏVE ORDERS

The simulation presented in chapter ‘4. Impact of High-Frequency Trading on Volatility’ showed how a market under stress is at risk of abnormal volatility due to the latency experienced by Low-

Frequency traders. The time a slow trader's market order takes to reach the exchange server and to execute may lead to its order being executed at a different price from the one originally intended. This may occur in both directions: while a market order wanders around the network, price may move either in favour or against the slow trader who posted it. Commercially available market data of course do not display the originally intended price of a market order. Therefore, without more detailed data it is impossible to appreciate this phenomenon. The situation is somehow different for limit orders. Let us consider a book with minimum bid-ask spread and a LF trader quoting a limit order at the top-of-book. If during the time this order takes to reach the exchange server the liquidity at the top disappears (either because consumed or cancelled), the limit order will still quote the intended price, the only difference being that the new quote will now sit in isolation at the top of the book. Also in this case it is not possible to decide whether the isolated order is due to latency or it was intended to quote a competitive price above the top-of-book. However, if the gap between the newly quoted limit order and the second best price is larger than one tick, it is reasonable to argue that the original intention was to quote an order at the top of the book or just above it. Quoting a limit order two or more ticks above the top-of-book is not a rational behaviour, as the same competitive advantage could be obtained by quoting just one tick above the top-of-book, with higher potential profit. These orders can be called naïve, as they appear as non-rational, even if the naivety may just be due to the latency suffered by the limit order. An example is a bid-ask book quoting, say, 100-105, where the tick is equal to 1. A competitive but still rational limit order might be a bid at 101 or an ask at 104. Any limit order (either bid or ask) at 102 or 103 can be considered naïve, since the same competitive position in the book could be obtained at a better price. Analysis of the order books for the periods investigated above on May 3rd through 7th, 2010, reveals that on the day of the Flash Crash there were 431 so-called naïve orders, against 5 on May 3rd and 4th, and

16 on the two other days. Table 25 displays the naïve limit orders quoted on May 6 between 18:39:00 and 18:45:28 GMT, where the first column shows the number of orders in the burst, the second column indicates the time at which the first order of the bursts arrived at the exchange server, and the third column shows the price gap (a positive gap indicates a bid order, that is, a naïve limit order whose price is higher than the second best order, and a negative one indicates an ask order with a price lower than the second best). Moreover, whereas on the other days the maximum gap between the top-of-book and the naïve orders was never larger than two ticks in either direction, on May 6 the maximum gap for the bid book was 4.75 index points (at 18:45:18.707 GMT), equivalent to 19 ticks, and 3.25 index points on the ask book (at 18:45:27.998 GMT), equivalent to 13 ticks. This suggests a very fast market when LF traders experienced a high delay in relative terms. A delay in absolute terms is measured in milliseconds, whereas a relative delay can be measured with respect to the speed of fast traders. This table is informative as it shows evidence of several anomalies occurring on May 6, 2010. A limit order posted with a gap of several ticks may be indicative of very high rate of price changes per unit of time, so that in the period between the price observation and the quoting of a naïve order, the market has moved up or down a lot. A naïve limit order is prone to be picked off by a trader sufficiently fast to exploit the opportunity, whereas a similarly naïve market order is likely to move the market abruptly, even without the conscious and explicit consent of the trader posting it. As stated above, commercially available data misses the information to decide about the reason behind the naivety, but the naïve order ratio between May 6 and the other days of the same week (86.2 on the first two days and 26.9 on the other two) suggests either a concentration of naïve traders on that day, or a frenetic HFT activity that caused slow traders to often accumulate large delays. Since it makes sense assuming that the delays experienced by limit orders were also affecting market orders, it is reasonable to state that the phenomenon

Time	Max gap	Time	Max gap	Time	Max gap	Time	Max gap	Time	Max gap	Time	Max gap
183921825	0.5	184434818	0.5	184503415	0.5	184512499	0.5	184518708	-1	184522311	1.25
183934242	-0.5	184435171	0.5	184503790	0.5	184512500	-0.5	184518708	-0.75	184522509	-1
184000955	0.5	184435328	-0.5	184504214	0.5	184512514	1.5	184518710	1.5	184522515	-0.75
184014735	0.5	184436587	-0.5	184506230	-0.5	184512516	-0.5	184518710	-0.5	184522773	-2.25
184025811	-0.5	184436922	-0.5	184506506	-0.5	184512523	0.75	184518710	-1.75	184522774	-2
184034992	-0.5	184437745	0.5	184506565	0.5	184512527	0.5	184518710	-1.5	184522780	-1.25
184036191	0.5	184437754	-0.5	184506902	0.75	184512527	1.5	184518710	-1.25	184522782	0.5
184132164	-0.5	184439203	0.5	184506902	0.5	184512531	-0.75	184518710	-1.25	184522797	-0.5
184135937	-0.5	184439230	-0.5	184506905	0.5	184512549	2	184518715	0.75	184522810	-0.75
184159458	0.75	184440942	0.5	184507202	-0.5	184512552	1.75	184518715	-1.25	184522830	0.5
184159462	0.5	184442992	-0.5	184508191	0.5	184512561	2.25	184518722	-1	184522834	-0.5
184215661	0.5	184442995	0.5	184508324	-0.5	184512561	1	184518722	-0.75	184522847	-0.5
184244102	-1	184443577	-0.5	184508589	-1	184512562	1	184518724	1.25	184522852	0.5
184244104	-0.5	184443586	0.5	184508589	-0.5	184512562	0.75	184518724	1.5	184522939	0.5
184248304	-1	184444065	-0.5	184508593	0.5	184512573	-0.75	184518727	1.25	184523414	-0.5
184248304	-0.75	184444431	0.5	184508715	0.75	184512575	-0.5	184518727	-1.5	184523578	-0.5
184248306	0.75	184444571	0.5	184508716	0.5	184512820	-0.5	184518728	1	184523669	0.5
184250483	-0.5	184444890	-0.5	184509599	0.5	184512868	0.5	184518729	1	184523897	-0.5
184300239	-0.5	184447324	-0.5	184509724	0.5	184512960	2.5	184518733	0.75	184524039	0.75
184300928	-0.5	184447676	0.75	184509769	-0.75	184512961	0.5	184518734	0.5	184524068	0.5
184307638	0.5	184447676	0.5	184509775	-0.5	184512963	1.5	184518734	0.75	184524138	-0.5
184308931	-0.5	184448270	0.5	184509776	0.5	184512965	-0.75	184518734	-0.5	184524789	-0.5
184311106	0.75	184448373	0.5	184510033	-0.75	184512971	-0.75	184518738	0.5	184525630	-0.5
184311106	-0.5	184448484	-0.75	184510034	-0.5	184512971	-1	184518738	0.75	184525770	0.5
184312358	0.5	184448493	0.5	184510177	-0.75	184512971	-0.75	184518738	0.5	184525998	-0.5
184319126	-0.5	184448554	0.75	184510184	-1	184512971	-0.5	184518741	0.75	184526206	-0.5
184319127	0.5	184448554	0.5	184510187	-0.5	184512974	1.5	184518764	-0.5	184526229	0.5
184321515	-1.25	184448556	0.75	184510254	0.5	184512982	1.75	184518767	-0.5	184526235	-0.5
184321516	-0.5	184448566	1	184510260	-0.5	184512982	1.25	184518771	0.5	184526301	0.5
184322074	0.5	184448566	0.5	184510456	-0.5	184512984	0.5	184518859	-0.5	184526441	0.5
184322186	0.5	184448566	0.5	184510982	1	184512986	0.5	184518867	0.75	184526525	0.5
184322929	0.5	184448569	-0.5	184510983	0.75	184512993	0.75	184518872	-0.5	184526525	-3.25
184322954	-0.5	184448736	-0.5	184510986	-0.5	184512993	1	184518883	-0.5	184526526	-3
184329606	0.5	184448783	-0.5	184511003	0.5	184512993	1.25	184518898	0.5	184526530	-2.75
184332611	-0.5	184449336	-0.5	184511050	0.5	184512993	0.5	184519220	-0.5	184526530	-2.5
184337250	0.5	184449377	0.5	184511091	0.5	184513237	0.75	184519562	-0.5	184526530	-1.75
184337260	-0.5	184450202	0.5	184511102	0.5	184513502	0.5	184519639	0.75	184526531	-0.5
184340129	-0.75	184450209	-0.5	184511259	0.75	184513580	-0.75	184519663	-0.5	184526898	2.5
184340130	0.5	184450737	0.5	184511260	0.5	184513582	-1	184519688	1	184526902	-1
184340863	-0.75	184451222	0.5	184511365	0.75	184513584	0.75	184519688	0.5	184526905	1.75
184340864	-0.5	184452502	0.75	184511372	0.5	184513585	-0.5	184519699	0.5	184526911	2.25
184341150	0.75	184454087	-0.5	184511382	0.5	184513612	0.5	184519699	0.75	184526913	-0.5
184341166	0.5	184454127	-0.75	184511585	-1.5	184513882	1.75	184519702	0.5	184526913	-0.5
184343487	-1.25	184455052	-1	184511588	0.75	184513882	1.5	184519702	-0.5	184526950	1
184343489	-1	184455062	-0.75	184511588	0.5	184513882	1.5	184519831	-0.5	184526972	-1
184343613	0.5	184455066	-0.5	184511589	-0.5	184513882	1.5	184519851	0.5	184526987	1.5
184345146	-0.75	184455166	-0.75	184511597	0.5	184513882	1.75	184519963	0.5	184526992	-1.5
184345350	-0.5	184455169	-0.5	184511709	1.75	184513884	0.5	184520000	0.75	184526992	-1.25
184346987	-0.5	184456025	0.5	184511716	-0.5	184513887	-0.5	184520154	1	184527014	-0.5
184349267	0.5	184456105	-0.5	184511746	0.5	184514550	0.5	184520158	0.5	184527016	1.75
184353775	-0.5	184456357	-0.5	184511748	-1.75	184515006	0.5	184520946	-0.5	184527016	-1
184353777	0.5	184457372	-0.75	184511751	1.75	184516213	0.5	184521103	0.5	184527017	1.5
184402494	-0.5	184457779	0.5	184511751	0.5	184516653	0.75	184521103	0.5	184527017	0.5
184405333	-0.75	184457894	0.5	184511760	1	184516654	0.5	184521120	-0.5	184527018	1
184405333	-0.5	184457984	0.75	184511785	0.5	184517961	0.75	184521133	-0.5	184527018	-0.5
184406276	0.75	184458533	1.25	184511785	0.5	184517961	0.5	184521358	-0.5	184527019	1.25
184406276	0.5	184458541	0.75	184511786	0.5	184517963	2	184521418	0.75	184527020	1
184406327	0.5	184458541	0.5	184511799	1	184517966	0.5	184521583	0.75	184527020	-0.75
184417508	0.5	184459537	0.5	184512185	-0.5	184517969	0.5	184521583	0.5	184527024	-0.5
184417516	-0.5	184501782	-0.75	184512466	1	184517974	-1	184521585	0.75	184527028	-0.5
184427912	0.5	184501782	-0.5	184512466	2	184517975	-1.25	184521585	0.5	184527039	-1
184428061	-0.5	184503005	0.5	184512469	1.5	184517975	-1	184521631	1.25	184527043	0.5
184428072	0.5	184503171	0.5	184512471	0.5	184517975	-0.75	184521634	1	184527065	-0.5
184428072	0.75	184503187	-0.5	184512489	1	184517975	-0.5	184521634	0.75	184527093	-0.5
184428077	-0.5	184503197	0.75	184512489	1	184517980	0.5	184521672	0.5	184527168	-0.5
184428078	0.5	184503197	0.5	184512491	1.5	184517993	-0.5	184521863	0.5	184527198	-0.5
184428086	-0.5	184503389	-0.75	184512491	1.5	184518541	-0.5	184522001	0.5	184527325	0.5
184428125	0.5	184503396	-0.5	184512491	1.5	184518585	0.5	184522017	0.75	184527960	-0.5
184428180	-0.5	184503411	1	184512491	1.5	184518609	-0.5	184522150	0.5	184527997	1.75
184429888	-0.5	184503412	-1	184512491	-1.25	184518707	4.75	184522157	0.75	184527998	-3.25
184430377	-0.5	184503412	-0.5	184512494	1.25	184518707	4.5	184522174	-0.5	184527998	-3
184430507	-0.5	184503412	-0.5	184512498	0.75	184518708	4	184522252	-0.5		

Table 25. Naïve limit orders during the Flash Crash

described in chapter 4 did actually materialise. It must be clear that the number of naïve limit orders in itself carries no much significance but it is a proxy of the number of naïve market order, which likely contributed to move the market abnormally for the reasons explained in chapter 4, together with the Stop Loss orders, as discussed in this chapter.

7.3.14. DISCUSSION

The main result that can be drawn from the previous observations is that the Flash Crash on the E-mini S&P 500 futures was, at some considerable extent, caused by a combination of falling prices, Stop Loss orders, and decreasing liquidity, all glued together by HFT. As each of these three factors was the consequence of the previous and the cause of the following one (falling prices trigger Stop Loss orders, which consume liquidity, causing a further fall of prices, and so on), a devilish feedback loop ensued. It is important to notice that each of these conditions can, and does, often appear on its own in the market without necessarily resulting in a memorably negative day, and even two conditions at the same time can co-exist without leading to a crisis. When all three conditions turn up in a high-frequency environment, as on May 6, 2010, chances are higher that a local crisis does materialise. The findings shown above are compatible with CFTC-SEC (2010a), that recognises the heavy use of Stop Loss orders leading to reduced liquidity as a promising path of investigation. These findings also match the opinion, stated in many important articles [for example Menkveld and Yueshen (2016)], that the large selling programme launched by Waddell & Reed was not the main factor in the crash. Leland (2011) and Zigrand, Cliff and Hendershott (2012) acknowledge Stop Loss orders, together with forced margin sales, as one of its main causes. Those, and other studies, highlighted the contribution of several factors together as the causes of the Flash Crash but in a more general way than shown in this chapter. This research precisely identifies in those three factors and in the mechanism linking them together a plausible and likely cause of the

Crash. It also provides a quantitative analysis of its findings. Moreover, the section about naïve orders confirms, albeit indirectly, the mechanism of price volatility suggested by the simulation experiment carried out in chapter 4. In all cases in which it was possible to compare theoretical (Petri Nets) and empirical (simulation) results with real data, the latter provided a good matching.

In this analysis, as in the simulations presented in chapter 4 and chapter 5, the entire financial horizon was restricted to only one market, with the implicit consequence that a local crisis could not propagate across market boundaries. It is admittedly a limit of this research; a more realistic testing environment should be set up and used as an 'economic wind tunnel' for ex-ante crisis prevention or, at least, for ex-post crisis investigation purposes, as suggested by Sornette and van der Becke (2012). Having recognised the importance of a multi-market simulation, it must be said that even a one-market model can shed some light on the most dramatic financial event of this decade, so far. Obviously, not all three factors carry the same weight and play the same function in a crisis. Out of the three simultaneous conditions identified as the main contributors to the Flash Crash, Stop Loss orders are the most frequent in day-to-day operations; they are common practice and no financial authority is worried about them. Nevertheless, they can be important contributors, especially when many of them pile up, ready to trigger as the price movements take a definitive direction. It is also intuitive that in a stressed market most investors would be cautious enough to protect their trading with Stop Loss orders, even against small swings. Sharply falling, as well as uprising, prices are occasional but not infrequent occurrences; they are intrinsic to market practices - and fortunately so: frozen markets are not desirable from any participant's point of view. Investors look for price dynamics and lack of it would make financial activities unappealing. Scarce liquidity is different: it is a major threat on itself. Regulators and exchanges are engaging a full-time struggle to ensure more and more abundant liquidity. High-Frequency Trading has found several supporters on the basis that this practice tends to increase market liquidity. Nevertheless, on May 6 liquidity virtually disappeared and lack of liquidity exacerbated the combined effect of the other two factors. However, even scarce liquidity on itself is not the automatic cause of a major crisis: if prices are stable the macro-effect would be scarcely noticeable, for example when prices move one tick up, then one tick down, and so on and so forth. There are securities, and even entire markets, frequently or permanently affected by scarce liquidity but they do not necessarily experience daily crises. Therefore, from all the previous considerations it sounds sensible to state that the non-linear input-output transformation effect (aka 'butterfly effect') was the real cause of the Flash Crash. Had the market had the capability to prevent an apparently innocuous cause to turn into a violent outcome, to avoid prudential withdrawals from a downward market leading to a crash, or were a bridge designed to prevent the turning of a breeze from developing into an unbearable mechanical oscillation, some of the disasters mentioned above would not have happened. The butterfly effect seems to be at the root of most critical problems the financial markets (and, perhaps, the Mankind as a whole) are currently facing. Systems have apparently grown too complex and too rapidly for systems theory to be able to cope with.

7.3.15. CONCLUSION

It turns out that beyond all the financial stability aspects listed in the literature review, resilience seems to be an indispensable feature markets must possess in order to ensure an orderly functioning. Resilience is a term originally used in engineering but recently it has also been widely used in other fields, ecology and social sciences among them. A system is resilient if it has the capability to absorb the effects of a disturbance while retaining the same features as far as output, structure, quality, and response time are concerned. A study from the Bank of England [Anderson et al. (2015)] recognises that "[r]esilient markets provide predictable access and liquidity for funding, investing, saving and risk transfer, and are underpinned by robust infrastructure. Market liquidity contributes to Resilient markets by ensuring that changes in prices are orderly and largely reflect changes in valuation factors such as the outlook for future cash flows" (ibid. p.15). The same paper defines the concepts of market 'amplifiers' and 'stabilisers'. "Amplifiers are market dynamics that act to reinforce buying or selling pressure in response to an initial price move" (ibid. p.16), while stabilisers act in the opposite direction. Both dynamics depend on the market structure and the nature, preferences, goals, investment horizons, beliefs, and strategies of the participants. In particular, homogeneity in these factors is likely to reinforce the amplifying effect (if all investors are bearish about the market, they are all likely to sell, amplifying an initial price movement - and conversely for a bubble in case of bullish sentiment), whereas diversity plays a stabilising role. A financial system can be said resilient if it encompasses mechanisms that swiftly bring it back to equilibrium, that is, in which stabilisers dominate amplifiers, whenever a disturbance arises. A disturbance can be a price misalignment, so in this respect HFT, with its arbitraging capability and despite the issues discussed in the previous chapter, is definitely a factor increasing the resilience of the system. HF traders, as non-registered market makers, are strong contributors to liquidity supply and the literature confirms that they are more often supplying liquidity by quoting limit orders rather than aggressively taking liquidity away from the market. Volatility is another hot topic in trading and once again, albeit with some very noticeable and not infrequent exceptions like troubled markets, HFT seems to be a mitigating factor in this respect, as fast traders are also quick to close positions as soon as a minimal profit can be extracted from a price swing, in so doing driving prices back to the mean. All these effects are intrinsic to market behaviour, usually very well tolerated, and often sought after, by the system. The greatest insurance markets have to guarantee an orderly functioning is provided by the large amount of participants - and the differences among them. Their number facilitates mean price recovery in case of misalignment, return to equilibrium when volatility is excessive and supply of liquidity on a continuous basis. Volatility needs not to be directly linked to illiquidity. A market may display abundant liquidity but perhaps at a different price from the one traded last: liquidity is there but someone will have to accept a loss to make use of it, so prices will move. The opposite case is more troublesome: if liquidity is low even moderate level of trading has the potential to make prices volatile. However, this is not an uncommon occurrence: no market can be guaranteed to be highly liquid at all times. But usually this is a temporary condition; sooner rather than later other participants, noticing the scarcity of liquidity, will judge it profitable to supply limit orders, perhaps at a different price from the one traded last, restoring the normal functionality of the market. As long as there are plenty of players, dis-homogeneity is more likely than homogeneity, and the amplification risk seems to be easy to keep under control. Yet, something different happened on May 6, 2010. In the few minutes labelled as 'Flash Crash' volatility peaked, liquidity virtually disappeared, and the number of participants dried up. At that point all the resilience mechanisms usually relied upon for restoring 'normality' (whatever that might mean) failed miserably with results that are well-known. That event made clear, if necessary at all, that the financial markets follow a M-shaped equilibrium function: for small disturbances the stabilising forces are predominant and they tend to restore equilibrium (the V-shaped part of the profile, or the inner part of the 'M'). Yet, when the disturbance hits a certain threshold, the graphical description reaches the maximum of the M-shaped profile and an unstable condition is generated, not dissimilar from a heart subject to fibrillation. Over that threshold the resilience of the system collapses and the amplifier mechanisms overcome the stabilisers. This abnormal situation can only be cured by a dramatic intervention, like a defibrillating device - or in the Flash Crash case, by the CME Globex Stop Logic mechanism. The Chaos theory has studied a lot about the behaviour of stable-to-unstable transition and more use of it in the field of finance is auspicable to better understand complex market dynamics.

7.4. THE IMPACT OF ABSOLUTE SPEED ON MARKET STABILITY

7.4.1. INTRODUCTION

Most academic studies dealing with speed of trading highlight as the very core of the matter being the relative speed of HF traders with respect to slower ones. This is certainly the most striking factor for the reasons explained so far. However, there is another aspect only marginally taken into account by the literature: the speed of the trading client versus the speed of the exchange server. Quite characteristically, when Farmer and Skouras (2012a), Friederich and Payne (2012), Moallemi and Sağlam (2013), Hanif (2012), and Vuorenmaa and Wang (2014) address the topic of absolute speed, they always do so for contrasting it against relative speed and to assess the latter as important and the former as unimportant. The next sections discuss order cancellations after a run, to assess the weight carried by the speed of the exchange and its impact on trading in presence of cancelled orders. In this sense the speed of the exchange can be regarded as 'absolute', given that it is an invariant component of the system.

7.4.2. NUMBER OF CANCELLATIONS AFTER A RUN

An interesting topic to observe, linked to the considerations about absolute vs. relative speed, is the number of cancellations executed after a run. Their importance refers directly to the CME Globex rules (other exchanges may behave differently). One important safety feature implemented in the Globex platform is the pre-computation of the effects of a trade. In other words, before each trade is carried out, the platform simulates what would happen, and only if the outcome lies within the allowed range the execution is given the green light. In particular, the price delta following the sequence of Stop Loss orders triggered after a trade is computed before executing the trade. If the computed price moves more than a pre-determined amount up or down, then before executing the order a Stop Logic mechanism is launched, which pauses trading yet allowing quoting orders. This implies that the book situation is frozen at the moment the simulation is started. If an investor wishes to cancel an order while the exchange server is busy in simulating the consequences of a trade, its order is queued for execution after the current trade, and all the related Stop Loss orders, are carried out. This means that once a trade is possible according to the rules of the exchange, there is no way to exploit its latency to cancel an executable order, even if deep in the book. This prevents a potential anomaly described later. Nevertheless, investors worried by the large price movements which occurred on May 6, 2010 were probably not wasting time in simulating the status of the market after execution (even because they were not aware of the Stop Loss orders waiting behind the scenes) but tried to cancel stale orders before they were executed at a price likely to be penalised by market movements. The frustrated attempts to cancel quotes already on their way to be matched with Stop Loss orders are not displayed in the commercially available audit trail data. Yet, cancellations referring to orders outside the trading sequence did actually go ahead. As an example, let us suppose there is a bid book with three quotes: 100, 99 and 98, and a Stop Loss order ready to trigger as soon as the price reaches 99. When a large market order arrives, it consumes the liquidity at 100 and starts eroding the liquidity at 99. But a trade at 99 triggers a Stop Loss order at that price, which consumes all the remaining liquidity at 99. During this process, an investor that wished to cancel its bid at 99 would be denied the chance to do so because as soon as the trade at 100 starts, since all the orders affected by the sequence of Stop Loss orders are frozen at the beginning of the process. Yet, an investor may wish to cancel its bid quote at 98 and, if that order is still in the book after the platform terminates the trading run linked with the initial trade at 100, the quote gets actually cancelled. The number of cancellations at the top of the book after a trade run is a proxy of the market sentiment: the higher their number and the more nervous the markets. Moreover, this measure can also provide an indication of the cancellations not successfully executed because of their matching with active Stop Loss orders.

Panel C in table 12 shows the number of cancellations within N milliseconds after the termination of a run, with N taken equal to 10, 3 and 1, in order to consider various latencies of the system. In table 26 the row underneath each time window shows how much top-of-book orders cancelled on May 6 exceed the top-of-book cancelled on the day shown in the column itself.

	CANCELLATIONS				
Date	**03-May**	**04-May**	**05-May**	**06-May**	**07-May**
Within 10ms	11,690	9,537	9,885	15,542	11,106
	32.95%	62.97%	57.23%		39.94%
Within 3ms	4,830	3,994	4,533	8013	4,342
	65.90%	100.63%	76.77%		84.55%
Within 1ms	2,359	2,100	2,176	4125	2,059
	74,86%	96.43%	89.57%		100.34%

Table 26. Cancellations after a trade run

For example the percentage increase after 3 milliseconds on May 6 with respect to May 5 was 76.77%: (8013-4533)/4533=0.7677. It can be noticed immediately that the number of quotes cancelled within 1 millisecond after a run on May 6 is much larger than on any other day, ranging from nearly 75% on May 3 to more than 100% on May 7. Indeed, among all the measures, the 1ms-latency looks the most reasonable for the purpose. On the one side it shows the cancellations occurring immediately after the end of a run or even those occurring during a run and reported immediately thereafter. On the other side, the differences in percentage are highest among all latencies, which suggest that the cancellation-during-a-run effect, very visible in the immediacy after the run, was less so with the growing of the time window. However, whether or not this assumption holds, May 6 experienced a strong increase of this effect with respect to the other days.

The conclusion can be that the abnormal market behaviour observed during the Flash Crash led many HF traders to near-panicking, resulting in a considerable increase in their cancellation rate. (It make sense to state that computers may panic: behind any software there is a human programmer, and if the program dictates to quickly cancel orders when certain conditions materialise, the computer will duly obey. If the critical conditions materialise at a high rate, the frenetic order cancellation activity mimics a panicking behaviour). However, as seen in the literature review chapter, many (human and electronic) traders judged the situation far too uncertain and preferred instead to withdraw altogether.

7.4.3. DOES IT MAKE SENSE TO TALK ABOUT ABSOLUTE SPEED IN TRADING?

In the literature and in the simulations presented in the previous chapters, the real meaning of speed was relative to an investor compared to the others. That is the main topic around which all the HFT debate goes. There were exceptions, though. In physics, the relativity theory states that all speeds are relative, with the only exception of the speed of light, which is absolute. The speed of light is the reference for all the measures of the system, which in that case is the Universe. In a financial system made up of an exchange and several traders, the reference speed of the system is the speed of the exchange server (in a multi-venue environment the reference speed is a complex function of the speed of all the exchanges). In the simple case of one exchange system the exchange server's speed is 'absolute' in as much it is the reference speed of the whole system, the difference with the relativity theory being that the exchange server's speed does not need to be the highest speed in the system. It is the reference speed of the system because it is the one which dictates the tempo to all other entities. This can be best explained with an example. Let us suppose an exchange server which has an internal clock and it performs one operation at each cycle of the clock, whatever its granularity. The quoting process is handled atomically by the system: it is read from the port at which it arrives and is written onto the book. The trading process is intrinsically slower, that is, it takes more than one clock's cycle. The process reads the price of the incoming market order, then reads the top of the appropriate book, then the two values are compared and if they match, then the trade is executed and the traded amount is subtracted from the appropriate top-of-book quantity. If the trading amount exceeds the quantity in the book, the next entry in the book must also be taken into consideration for supplying the liquidity needed for completing the trade. The transaction must also be recorded somewhere for the Clearing process to proceed with its own activities. Obviously, all the tasks just described cannot be distributed over different processors because the activities are inherently sequential and they access the same entity, namely the entry at the top of the bid or ask book. Concurrent writing on the same entity by more than one processor would jeopardise the consistency of the database, i.e. of the books. So, at least conceptually, the trading server needs to carry out more operations than the quoting servers. This is the reason for which all the major exchanges are continuously striving to increase their performance. On October 20, 2010 the London Stock Exchange issued a press release claiming an average trade time of 126 microseconds and 99.9% of trades completed in less than 400 microseconds, a performance which put its Turquoise trading platform at the top of the world. Other exchanges are competing and the contest resembles, once again, to the well-known arms race. Yet, despite of using the latest-latest technology, the fastest-fastest processors and so on, there is never the complete guarantee that the incoming order traffic would not clog the system, simply because the number of the investors, and their activity, may grow indefinitely. The following section replicates the scenario described in paragraph ‘7.2. A Simulation Using Petri Nets’, adding an incoming cancellation order during the execution of a trade. This apparently minor modification could pose an enormous conceptual problem which, under some extreme, but not unthinkable, circumstances, may become very real.

7.4.4. IMPACT OF EXCHANGE LATENCY USING PETRI NETS

With reference to figure 10, and assuming as usual both the bid and ask books empty, a 'bid SL' transition fires with bid price, as in the previous example, 87 and Stop Loss at 85. Other 'bid' transitions fire at 86, 85, and 84. Eventually, an 'ask' transition fires at 87. Transition 'match bid-ask' can fire and since the bid and ask price match, both the transitions 'bid SL OK' and 'ask OK' fire, setting all the conditions for a transition 'execute' to fire. Transition 'bin ask exec fires and the right-hand side of the Petri Net is finished. On the left-hand side, the bid order had a Stop Loss associated with it, the transition trigger bid SL' fires setting up the right conditions for the transition 'match bid-SL' to fire on its turn. In this case the trading price (87) and the Stop Loss price (85) are different, so they do not match. Transition 'bin bid-SL' fires and the transition 'bin price' cleans up the net. Then an ask order is posted at 86, which fires the transition 'match bid-ask' against the highest bid quote, at 86. The two prices match, causing 'bid OK' and 'ask OK' to fire, and therefore the transition 'execute' also to fire. The transition 'match bid SL' compares the trading price (86) and the Stop Loss price on hold (85) and since they do not match, 'bin bid SL' and 'bin price' fire, cleaning up the situation. Eventually an ask order at 85 arrives. Orders on both the bid and ask books exist and thus they trigger the transition 'match bid-ask', comparing the highest bid with the lowest ask. They match at 85 so transitions 'bid OK' and 'ask OK' fire, paving the road for an 'execute'. At this stage a change from the example in paragraph 7.2 occurs. The investor which posted the bid order at 84 notices the rapid price drop and gets worried that its order at 84 may soon become stale. Therefore it cancels the order at 84 and replaces it with another similar order, this time at 83. This is possible because the trading processor only locks the orders at the top of the book, allowing book changes underneath that do not affect the current trade. Then transitions 'bin bid exec' and 'bin ask exec' as before. The transition 'match bid SL' matches the trading price at 85 to the Stop Loss price, firing 'book ask SL', transforming the Stop Loss order into an ask market order ready to execute at best. Transitions 'bin price' cleans up. Now the order at 84 is no longer there; at its place there is the newly-posted limit order at 83 on the bid side and a market order (originated from the transformation of the bid-SL, on the ask side). So, despite the Stop Loss order was originally posted at 85, given that there is no limit order at 85, and that the limit order at 84 in the meantime has been replaced with a limit order at 83, when transition 'match bid-ask' eventually fires, the market order matches with the best limit order it can find - at 83. This time transition 'execute' fires at an even worse price than in the model described in paragraph 7.2. Obviously, this scenario can only occur in a situation of scarce liquidity. The example simplifies the situation, pretending that one order was able to consume all the liquidity at a certain price level, but as seen in the Flash Crash data analysis, such supposedly extreme situation do occur in real life. The final effect of the example just described is an even more rapid price decrease. A similar situation on May 6, 2010 favoured market makers to withdraw, liquidity to shrink further, prices to nose-dive, and panic to spread quickly. Unfortunately, no matter how fast is the technology used by the exchange, nothing can prevent a large enough critical number of investors to supply, to cancel, and to take liquidity away at a rate capable to overwhelm any network, any firmware, and any processor.

In a world of human traders, their implicit latency with respect to the exchange made the critical number of investors so high to be virtually impossible to reach. But when algorithms are able to operate one thousand, one million, or one billion times faster and more frequently than humans, the number of HF computers needed to reach the same critical level of market activity, is reduced by the same ratio.

7.4.5. CONCLUSION

As already noticed, the scenario just described did not happen on the E-mini S&P 500 futures market. That market implements a feature that, at the cost of performing several preventive operations before actually launching a trade execution, prevents an order in the trading flow to be cancelled before the current order execution completes. Nevertheless, even the Chicago Mercantile Exchange is vulnerable to the Stop Loss scenario described in paragraph 7.2. But if the CME is immune from the order cancellation scenario described in the previous section, other exchanges are not. Some of them do not bother to simulate the whole trading process before executing it, with the advantage of saving precious (absolute) machine time but running the risk of seeing liquidity to disappear under their eyes, via the cancel-while-trading mechanism described above. The paper by van Kervel (2015) explains exactly the issue of ghost liquidity - same liquidity quoted on several exchanges and quickly removed as soon as one order has been taken by an aggressive trader. This practice, according to the author and many others, is most common in HFT time. An in-depth research on this topic should definitely be carried out but, because of confidentiality agreements between the exchanges and their customers, the identity of the traders is generally not available to third parties and therefore it is not possible to make any assumption on whether an order has been posted or cancelled by a HF trader, or otherwise. An anonymous id, secretly but uniquely identifying each trader could be useful for this research, but even that piece of information is generally unavailable on a commercial basis. It is true that a few world-renowned scholars have occasionally been granted access to the identity of the traders but this possibility is not usually available to most mortals. In order to cope with this problem, a set of proxies must be used - and proxies provide useful although only orientative information. This was the approach used in the previous pages. All the approaches used in research have some pros and cons and the one adopted here, market proxy, is no exception. It delivered results but such results need to be scrutinised to assess their validity and robustness – something that has been discussed along with each simulation or analysis, and that will be summarised, put together and concluded in the next chapter.

8. CONCLUSION

8.1. INTRODUCTION

This research focused on if, and how, and why, High-Frequency Trading has an impact on some important aspects of financial stability, namely market volatility, market tiering, arbitrage, and the recent sudden financial crises, the Flash Crash being only the most striking, but by no means the only one.

Financial volatility is probably the topic most dealt with in the headlines. Investors, even casual ones, and not necessarily only practitioners, are used to hear about the stock market gaining or losing a few decimal points on a daily basis. They all know that small swings occur and nobody is particularly worried by that. But when the daily change approaches, or trespasses, some psychological threshold, for example one percent, volatility starts to become an issue.

Market tiering is far less known and usually not recognised as an issue by the general public and experts. Even academia has only recently suggested that such a possibility may actually exists. Yet, it marks an anomaly to the orderly functioning of the markets.

The matter of arbitrage is well-known by most investors and some of them, because of their trading strategies, are labelled as ‘arbitrageurs’. However, advocates of the Efficient Market Hypothesis struggle to accommodate arbitrage in their theory since one of the most basic assumptions about financial markets is it being an even playground for all participants. Minor exceptions are tolerated but a systematic deviation from that assumption risk to jeopardise public’s confidence in the fairness of the system.

The main topic addressed by this research is how and why anomalies or crises materialise in HFT era. The Flash Crash was only the tip of the iceberg and even the general press occasionally informs the public about occurrence of a sudden large price swing in a stock, a currency or a commodity, usually reverting rapidly. As long as the end-of-day closing price does not display dramatic difference to the closing of the previous day such events are still tolerated. Nevertheless a crisis is a crisis and its limited time span must not divert attention from the intrinsic threat any crisis may pose to the stability of the financial system.

The next four sections of this paragraph will summarise the findings and the considerations developed in the previous chapters (HFT impact on volatility, on market tiering, on arbitrage, and on the Flash Crash), while paragraph '8.2. Discussion' will address the overall matter, reaching a conclusion for this research. The last paragraph indicates the possible paths for future research on the impact HFT has, and will have, on financial stability.

8.1.1. IMPACT OF HFT ON MARKET VOLATILITY

Chapter 4 presented an Agent-Based Model for investigating the extent to which HFT influences volatility. The discriminant between the two cases under study was the general state of the market. Markets behave according to several endogenous and exogenous factors; among the latter ones, markets are certainly influenced by news. Therefore a market can be relatively quiet (the situation dealt with in the first case) or relatively volatile (as in the second case) because of external reasons. Moreover, both cases showed one scenario with Low-Frequency traders only (QR=0) and others with various ratios of HFT-to-LFT activity (QR>0). A quiet market did not show any significant difference in volatility, whether High-Frequency traders participated or not. Instead, a market following a price trend showed a significant and very strong impact of High-Frequency Trading on volatility. The reason can be explained by noticing that traditional players experience a delay (because, for example, receiving price information through institutional channels rather than directly from the exchanges, slower order generation, networking latency) before having their orders executed, whereas HF traders instead enjoy immediate execution. The delay LF traders experience may lead, when prices move rapidly, to market orders being executed at a price different from the one originally intended. This does not necessarily translate into a loss: if prices move favourably to the slow traders, they may even make unexpected gains. However, uncertainty of price execution carries the potential to exacerbate price movements, especially if prices are already volatile on their own. This result confirms the literature, for example Kirilenko et al. (2011), who find an exacerbating effect of HFT in volatile markets. Interestingly, as more HF traders enter the game, trading latencies tend to become homogeneous again (when all latencies are short), as it was in the slow traders only scenario (when latencies were all long). With homogeneous latencies, whether short or long, volatility returns to its usual, lower values, compared to the situation with dis-homogeneous latencies.

8.1.2. IMPACT OF HFT ON MARKET TIERING

In several cases the literature [Cartlidge and Cliff (2012) among others] suggested that HF traders show the tendency to mostly trade with their similar fellows. This looked like a sufficient reason for studying the possibility of market tiering in a HFT environment. The Agent-Based Model presented in chapter 5 showed that HF traders' speed and strategies (as described by the literature) may lead to some degree of market tiering. Evidence displayed by the simulation was quite strong in case of HF-to-HF trading and somehow weaker in case of LF-to-LF trading. Typical values for the former case ranged between 60% and 70% in case of thin bid-ask spread, whereas the latter case only showed between 40% and 50%, when the spread is large. Overall, it can be said that some kind of market tiering seems to materialise as HFT activity increases. This research confirmed the findings of Cartlidge and Cliff (2012) and Johnson et al. (2013), according to which the phenomenon appears as a smooth trend rather than sudden phase transition.

8.1.3. IMPACT OF HFT ON MARKET ARBITRAGE

The impact of HFT on arbitrage was studied by implementing a two-market simulation and forcing a price difference on the same securities between the two venues. As soon as this occurrence turns up, all traders rush to take profit of the risk-free opportunity and pocket an abnormal profit. However, because of their speed HF traders take the lion's share of such opportunities, leaving their slower counterparts with the near-to-nothing leftover. The simulation used real-world figures for the number of HF and LF participants, the ratio between their respective trading activity, and transaction costs imposed to the trading process. The result was significantly strong in favour of risk-free and consistent profits made only by HF trader. This showed a failure of the Efficient Market Hypothesis in presence of HFT activity and, depending on transaction costs, only profits five or six orders of magnitude lower for slow participants. The latter figure is still significantly different from zero but, as foreseen by Fama (1970), in presence of transaction costs, not significant from any practical point of view. Indeed, it would need between one hundred thousand and a few million times the number of arbitrage opportunities taken up by the average LF trader to make the same profit made by its HF colleagues. The reason the Efficient Market Hypothesis holds even in case of existence of arbitrage opportunities exploited by traditional traders, is that such opportunities are shared between all or very many traders, making the average profit very close to zero. This was stated in the Seventies of the 20th century and the discovery was awarded the Nobel Prize. But when, in the current days, a small number of ultrafast traders are able to exploit the large majority of arbitrage opportunities for themselves, the Hypothesis cannot hold any longer. Again, as it was the case for the impact on volatility, when the number of HF traders increases, the profitable opportunities will be shared among a much larger number of fast arbitrageurs, dramatically decreasing the profit-per-head, and leading the EMH to reaffirm its validity.

8.1.4. IMPACT OF HFT ON THE FLASH CRASH

So far this research showed the results of its simulations. This is not a negligible outcome: simulations can be assimilated to experiments in a laboratory, which used to be the main tool for many centuries of scientific research. Simulations present several advantages, like flexibility, possibility to study frontier cases, and usually affordable costs. Yet, the uncertainty of the simulated model is to be added to the uncertainty of the observation (of any observation). Moreover, without a comparison with the real world, the results of an experiment risk to remain numbers on a sheet of paper. Therefore, the main purpose of chapter 7 was to show compatibility between the theoretical results reached and the audit trail data available. The chapter presented a mathematical model, using Petri Nets, of an exchange, and demonstrated, in theory, the potential adverse impact of the Stop Loss mechanism on volatility. Data analysis was affected by some degree of uncertainty, because of incompleteness of the data. In particular, the data used for the analysis missed the trader's identity, making it impossible to discriminate between HF and LF activity. Indication of whether or not a trade was originated by a Stop Loss order was also missing. Thus, some phenomena had to be observed indirectly and proxies had to be used. Having taken these caveats into account, the data showed that Stop Loss orders were very likely to have exacerbated the Flash Crash. Moreover, the data on that critical day displayed an abnormally high number of so-called naïve orders, that is, those limit orders posted well over the top of the book, apparently with no rational reason. That was a sign that some limit orders were experiencing delay while the market was moving: when the order hit the book the originally intended price was no longer at the top (or just above it) but much over it. The large difference of naïve limit orders between the day of the Flash Crash and other days in the same week, suggests that a similar unbalance may have occurred for market orders, leading to exacerbation of volatility, as found by the simulation about the impact of HFT on volatility in chapter 4. A third finding of the data analysis exercise was a confirmation of the impact fast and frequent order cancellation may have on a rapidly moving market. The Petri Nets model showed how a volatile market might be further stressed by liquidity suddenly disappearing because of order cancellations. The abnormally high number of cancelled orders found in the data on the day of the Flash Crash confirmed the findings of the theory.

8.2. DISCUSSION

8.2.1. IMPACT OF HFT ON MARKET STABILITY

The results of all the simulations and the data analysis presented in the previous pages go in the direction of a clearly negative impact of HFT on market stability.

The simulation on the impact of HFT on market volatility (chapter 4) confirms that when markets are under stress, HFT may cause volatility to significantly increase. It must be noticed that, although the findings in the literature in this respect are varied, most authors tend to highlight the beneficial effects of HFT on volatility, instead. The results of simulation in chapter 4 have been confirmed by the analysis of naïve orders in chapter 7. Naïve orders (or similar concepts) have never been dealt with by the current literature. Neither has the concept of ‘absolute speed’, with that intending the speed of the exchange server, which may not be able to cope with the large amount of activity by ultra-fast participants (the only noticeable exception being the non-academic Nanex research). If the server is overwhelmed by too large a number of incoming orders, the book may change before a market order spanning more than one book level is completed, leading to deals occurring at prices different from the intended one. Stop Loss orders have been identified by some authors as a factor potentially exacerbating volatility. Yet, so far no study has attempted to quantify the phenomenon, nor analysed audit trail data in search for malicious Stop Loss order effects. However, an even more interesting result (never mentioned in the literature) is probably that when the percentage of HFT over all trading increases further, volatility seems to decrease, until it reverts to the same level as when no HFT occurs at all. The interpretation could be that as HF traffic grows, either because the speed of HF traders increases further or, more likely, because their number grows, most trading occurs between homogeneous participants, and the impact of latency fades away. The latency effect in the simulation has been implemented as a circular queue, to mimic the delay LF traders are subject to. But if most traders operate at high frequency, then no or very few orders get queued up and the impact of delayed orders or trades will no longer be significant. The effect on volatility of a market with (nearly) only HF traders would be similar to the one with only low-frequency participants. So, it is not the speed of HF traders that leads to abnormal volatility, but the mutual interaction of few fast and many slow traders. When the market returns to homogeneity, such pernicious effect will no longer display itself. This does not mean that just sit and wait for traditional operators to adopt HF technology and strategies will do the job. Recent results [Baron, Brogaard and Kirilenko (2012), Serbera and Paumard (2016)] of reduced HFT profitability may suggest that the HFT market has already saturated to a certain extent and further entries may not follow as expected. Indeed, as the number of HF traders increases, they will have to share the profitable opportunities made possible by ultra-high trading speed. This is a point to highlight: it is not HFT per se that causes volatility to peak but the mutual interaction between traders with largely different speed that does, where each category may well fingerpoint the presence of the other one as the main culprit.

The market tiering simulation (chapter 5) provides an original research on this topic. Very few previous studies mention the possibility that HF traders mostly deal within among themselves, and none provided a quantitative assessment thereof. The quantitative results of the simulation suggest that HF traders have some tendency to deal with each other, and that causes LF traders also to deal with other slow traders, albeit at a lesser extent than found in the HF-to-HF trading. This is also a kind of market instability, as it goes against the sought after behaviour of the system: markets should be open places where any participant is able to trade with any other, regardless of their relative speed. If one category is able to strike all of the deals regarded as less risky (that is, those for which the bid-ask spread is tighter), this could be interpreted as a distortion of orderly market operations. As we have seen, it is not the case that ‘all’ narrow-spread deals are struck by HF traders with other HF traders, but the results show a certain tendency into that direction.

A similar argument applies to the findings about the impact of HFT on arbitrage (chapter 6). Although with some exceptions, markets used to be generally regarded as efficient. Despite arbitraging was considered a matter of facts, the competition to reap its benefits was open to many participants. Therefore, it was usually deemed insufficient to make a living on its own. HFT may have changed the scenario, potentially allowing a few fast traders to grab most or all the arbitrage-led profits. Some authors argued that HFT could reap the lion’s share of arbitraging but no quantitative evidence is provided by the current literature. Again, this phenomenon may disappear as more HF traders enter the competition, or it may be the case that the ones currently in are already seeing their profits declining. This somehow confirms the results of abnormal risk-free profit only holding under the two conditions set in the simulation: large speed difference between traders, and only a small number of traders enjoying it. When either condition does not hold, the arbitrage opportunities get shared between many more participants and the abnormal profits become negligible or null, as stated by the EMH. (This also confirms the value of simulations, which can represent market conditions difficult to find in the real world). However, at the moment the whole matter of arbitraging seems a dispute between few sub-second traders, with no gain for the rest. Again, the appreciation of the general public of this uneven playground may be regarded as a threat to financial stability. The above does not necessarily mean that HF traders will make all the profits in the market, leaving the others with nothing to pocket in. Fundamental analysis, by its own nature, watches at the medium- to long-term and if the analysis is correct, slow but hard-working participants will likely reap the benefits. Also Technical Analysis, albeit sometimes with a shorter time horizon, has, according to its followers, yielded profits to those who could rightly interpret its signals. The real issue at a stake is the confidence investors may have in an instrument (the market) that no longer satisfies the purpose it was originally thought for. According to economic theory the market should be a fair place where resources flow towards the most promising companies, projects, and ideas. Instead, if it is regarded as an elitist circle where profitable deals are struck among a restricted community of fast but risk-repelling participants, leaving all the risk borne with by those external to that community, general public's confidence may vanish, with unpredictable consequences.

The findings of chapter 7, to a certain extent, confirm the outcome of the simulation presented in chapter 4. Moreover, both a mathematical model and the data analysis suggest that the combination of three factors (volatility, scarce liquidity and Stop Loss mechanism) in a HFT environment, may lead to disastrous consequences. The analysis of audit trail data supports this view. Frequent order churning, a peculiar characteristic of HFT, also might, under certain conditions, exacerbate volatility further, and once again the outcome of the theory finds (albeit indirect) evidence in the market data.

Overall, the findings of this research warn of HFT showing some highly controversial aspects, which understandably alert traditional investors [as noticed by Beunza, Millo and Pardo-Guerra (2012)], even though academic researchers are much less impressed. However, regulators and policy-makers would do better to carefully monitor the next future developments of High-Frequency Trading and of the low-latency arms race, ready to adopt bold measures as soon as the level of alert shall turn more concrete. All that, in the interest of orderly functioning of the markets and market stability in general.

8.2.2. POLICY IMPLICATIONS

As a matter of principle, any result showing an impact of High-Frequency Trading on financial stability should alert regulators and policy-makers. However, as stated in the previous paragraph, although the results of the simulations find some confirmation in the audit trail data analysis, such confirmation is only indirect. In presence of HFT, the literature mentions some signs of market tiering and inefficiencies. More debated is the issue of HFT-induced increase in volatility, although the simulation presented in chapter 4 is quite one-sided in affirming so. Moreover, the data analysis carried out in chapter 7 provides indirect evidence of such phenomenon. This is not a proper demonstration. However, two different methodologies, a simulation and real-life data analysis, both point in the same direction. The results have to be correctly interpreted, though.

(i) The simulation in chapter 4 states that in presence of a small number of HF traders in an environment populated by a large majority of slow traders, the terrain is ready, under certain conditions, for a possible sharp increase in volatility. It does not state that it did or will happen – simply that the potential exists for a crash mechanism to trigger.

(ii) The first Petri Nets-supported theory in chapter 7 shows a Stop-Loss scenario, compatible with HFT activity that can exacerbate volatility. This effect is made even more dramatic by ultra-fast order cancellations, as shown by the second Petri-Nets scenario at the end of the same chapter.

(iii) The data analysis in chapter 7 states that on the day of the Flash Crash the trade runs were abnormally frequent and, above all, that the price changes experienced within the same trade run were up to 13 times (3.25 index points versus 0.25) higher on May 6 than on the previous three days.

(iv) Also, the data analysis states that the number of cancellations in the millisecond after a trade run on May 6 was between 75% and 100% higher than on the other days of the same week.

None of the results shown above can be considered definitive. Lacking the identity of the traders responsible for the abnormal behaviour makes the conclusions indirect. However, the matching between the simulations and the results indirectly obtained by analysing the data provides sufficient reasons to raise worries about the state of financial stability in current times, especially in presence of several mini-flash crashes, as identified by a few scholars and practitioners.

Regulators and policy-makers need to work hard to prevent bursts of financial instability and, as stated by Danielsson (2013), they may actually do a very good job on the large majority of the times, since their successes are scarcely noticeable. For that reason, it is important to keep proceeding on prevention activities. The results of this research suggest a few aspects of the relationship between HFT and financial stability worth monitoring closely, in the attempt of minimising some likely scenarios of financial crises.

8.3. PATHS FOR FUTURE RESEARCH

Any research effort, and this research is no exception, could go further and its outcomes reach a higher degree of validity. In particular two aspects are natural candidates for a future research on the topic of HFT and financial stability: one operational and one methodological. The first aspect worth attracting further work, although needing more resources, is that both the simulations in chapter 4 and in chapter 5 mimic a one-market environment, with no mutual interactions with other venues. Although this assumption holds in the specific case of the E-mini S&P 500 futures contracts only traded at the Chicago Mercantile Exchange, the scenario could be investigated further. In fact, most markets are tightly intertwined and events tend to spread quickly across walls of institutions and borders of nations. The idea behind both simulations presented in this research was to demonstrate that, under certain conditions, abnormal behaviours may show up. If a volatility peak or a split market could exist, and the theory foresees it, being aware of such possibility is certainly a useful input. However, simulations that mimic a few venues operating in parallel and interacting with each other would definitely be a useful path of further research.

The simulation methodology, although insightful, can never substitute real life. In order to analyse whether the outcome of a simulation taking HFT into account does materialise when people's money is at a stake, there is no better way that looking at what happened in the real world, as reflected in audit trail data. The main difficulty is that the trader's identity is exchange-confidential and as such it is very difficult to spot who quoted a specific order or initiated a specific trade, even only if they were High- or Low-Frequency traders. Under certain conditions exchanges seem to disclose such information but the process to achieve that goal is not an easy one, nor easily successful. However, renowned research institutions may succeed where individual researchers failed. Certainly having access to confidential information about the category (HF or LF) the participants of a trading belong would greatly improve the reliability of the analysis. Being sure of how HF traders behave would clarify many aspects that, lacking inside information, can only be best-guessed, with a non-negligible percentage of possibility of error. Having a clear view of the market and its nuances would possibly provide evidence to pave the way to the missing knowledge in the field of High-Frequency Trading and its relationship with financial stability.

APPENDICES

APPENDIX A: PSEUDO-CODE FOR SIMULATION IN CHAPTER 4

The sequence of subroutine calls for the simulation of the impact of High-Frequency Trading on volatility is as outlined by the following pseudo-code.

```
Subroutine Launch{
  repeat 100 times{ //sufficient for hypothesis testing
    call Simulation subroutine
    call AnalyseTrades subroutine
  }
}

Subroutine Simulation{
  call InitialiseData subroutine
  repeat 10,000 times{
    compute spread as tick multiple
    select book type (bid vs. ask)
    select order type (limit vs. market) according to case
    select trader type (HFT vs. LFT) according to case
    if trader type is LFT then{
      add order data to circular queue
    else // HFT order
      if limit order then{
        if liquidity < max_liquidity then{
          call IncreaseLiquidity subroutine
        else
          if spread > 1-tick then{//next price level
            call InsertPrice subroutine
          }
        }
      else // market order
        call Trade subroutine
      }
    }
    if time at top of circular queue has expired then{
      if limit order then{
        if order inside spread then{
          if spread > 1-tick then{//next price level
            call InsertPrice subroutine
          }
        else // order deep in the book
          compute book row at which to operate
          if no orders at that price then{
                call InsertPrice
            elseif liquidity < max_liquidity then{
                call IncreaseLiquidity subroutine
            }
            else // market order
                call Trade subroutine
            }
        }
    }
}

subroutine IncreaseLiquidity{
    add trader's order at given price level
    increase liquidity counter
    record event onto logbook
}

subroutine InsertPrice{
    insert new price level above current one
    add trader's order at the new price level
    set liquidity at the new price level equal to 1
    record event onto logbook
}

subroutine Trade{
    cancel quote to be traded
    decrease liquidity by 1
    if liquidity is equal to 0 then{
        remove price level from book
    }
    record event onto logbook
}

subroutine AnalyseTrades{
    repeat for all events recorded in logbook{
        if event equal to TRADE then{
            if traders pair equal to LFT-LFT then{
                increase LFT-to-LFT counter
            elseif traders pair equal to LFT-HFT then
                increase LFT-to-HFT counter
            elseif traders pair equal to HFT-LFT then
                increase HFT-to-LFT counter
            elseif traders pair equal to HFT-HFT then
                increase HFT-to-HFT counter
            }
            update minimum price and maximum price
```

```
80 }
81 }
82 write all counters onto database
83 }
```

A.1 DESCRIPTION OF THE ALGORITHM

In the simulation two different cases were tested: the unperturbed (base) case and the trend case. The 'trend' case identifies a situation in which most market orders go into the same direction (either buy or sell) whereas the in 'base' case all parameters are randomly chosen. Each of the 100 repetitions (line 2) goes on for 10,000 cycles (line 9), each representing one activity on the market (limit order or market order). A trend starts at the first cycle and goes on for 2,500 cycles, then it returns to the unperturbed state until cycle 5,000; the trend starts again on the next cycle, to terminate at cycle 7,500. The algorithm works out the number of ticks the spread is made of (line 10) and it randomly selects the book upon which it will act (line 11). The order type (limit or market) is computed randomly (line 12) but, if the spread is wider than one tick, a higher probability is assigned to limit orders, in proportion to the spread. This matches common experience, according to which a wide bid-ask spread makes all traders more cautious, whereas a thin spread signals generalised consensus on the quoted price and makes market orders more likely. During a trend, if the randomly selected book is the opposite of the trend direction (set at the beginning of that run), then only a limit order is possible in that direction. The only restriction in place outside a trend is that no two consecutive market orders can be posted in opposite directions to each other: at least a limit order needs to be quoted in-between. This is to avoid a frenzy market with buy and sell trade orders with little underlying logic, which is against common experience. If the spread is thin, the trader is randomly selected according to the ratio established (line 13). If the spread is wider than one tick and a market order has been selected, then low-frequency traders are given higher priority than their HF counterparts, highlighting how HF traders feel more comfortable in tight markets. At the end of the 10,000 cycles, the difference between the highest and the lowest price, an indicator of volatility, is recorded (line 79). The algorithm is launched 100 times (line 2) to provide a suitably large amount of data, and the average and standard deviation are worked out over the number of launches. HF-to-LF quoting ratio equal to 0 (for the no HF traders scenario), 1, 2, 20 and 200 were tried, with the goal of finding a resulting trading ratio between 50% and 60% In both the base case and the trend case, the 1-to-1 quoting ratio (QR=1) yielded an acceptable HF-to-LF trading ratio but the other ratios were also useful to verify the change in volatility as this ratio varies. Moreover, in the simulations a feature called 'blinking' was also introduced. It means that a slow trader introduces a certain delay (a blink), in its trading process. A typical example is a human trader who needs a certain amount of time to realise how the market is moving, to make a decision, and to physically carry it out. A different possible source of latency is given by an algorithm which conveys its orders via a communications channel relatively slower than those used by HF traders. The duration of this delay may be regarded as blinking and it has been set to 650 times the elementary time unit. This delay is a configurable information being read at the beginning of the execution and could be modified in subsequent launches of the simulation. The choice of the number made for running the simulation is consistent with the minimum market cycle to be assumed equal to one millisecond and a very fast, yet complex, human activity estimated to last no less than 650 milliseconds. This means that when a LF trader is selected by the algorithm, its decisions are recorded into a circular queue (line 15) for later execution (line 29), simulating the delay due to human reactions or slow communications channels. After 650 time cycles, that order is actually executed. If it is a limit order and the liquidity is not at its maximum, it is posted at the originally intended price. In case the price is deep in the book, it is added to the existing liquidity (line 40). If the quote is above the current best bid or below the current best ask price, it is inserted at that price (line 33). If instead it was a market order, it is executed 'at best' (line 43).

APPENDIX B: PSEUDO-CODE FOR SIMULATION IN CHAPTER 5

The sequence of subroutine calls for the simulation of whether High-Frequency Trading creates a two-tier market is as outlined by the following pseudo-code (comments and explanations within the pseudo-code are preceded by a double slash '//').

```
subroutine Launch{
    repeat 100 times { //large enough number for hypothesis testing
        call Simulation subroutine
        call AnalyseTrades subroutine
    } // repeat
}

subroutine Simulation {
    call InitialiseData subroutine
    repeat 10,000 times {
        randomly select book type (BID or ASK)
        if spread > 1 tick then {
            select trader (LF or HF) giving higher chances to LF
            randomly select order type (limit or market), with higher chances to limit orders for HF traders
        } // if spread > 1 tick
        else { // spread = 1 tick
            select trader according to quoting ratio (QR)
            randomly select order type; higher chances to market order if liquidity is deep
        } // if spread = 1 tick
        if order type = LIMIT {
            if liquidity < MAX_LIQUIDITY then {
                call IncreaseLiquidity subroutine
            else
                if spread > 1 tick then {
                    call InsertPrice subroutine
                } // if spread > 1 tick
            } // if liquidity < MAX_LIQUIDITY
    else // order type = MARKET
        call Trade subroutine
    } // if order type = LIMIT
} // repeat

subroutine InitialiseData{
    read no. of fast traders
    read no. of slow traders
    read ratio between fast and slow traders' orders
    initialise bid and ask books
    format output sheet
}

subroutine IncreaseLiquidity{
    add trader's order at given price level
    increase liquidity counter
    record event onto logbook
}

subroutine InsertPrice{
    insert new price level above current bid or below current ask
    add trader's order at the new price level
    set liquidity at the new price level equal to 1
    record event onto logbook
}

subroutine Trade{
    cancel quote to be traded
    decrease liquidity by 1
    if liquidity is equal to 0 then {
        remove price level from book
    }
```

```
55 record event onto logbook
56 }

57 subroutine AnalyseTrades{
58   repeat for all events recorded in logbook {
59     if spread = 1 then {
60       if event is a TRADE then{
61         increase the appropriate counter // LF-LF, LF-HF, HF-LF, HF-HF
62       } // if event is TRADE
63     else // spread > 1
64       if event is a TRADE then {
65         increase the appropriate counter // LF-LF, LF-HF, HF-LF, HF-HF
66       } // if event is a TRADE
67     } if spread = 1
68     if applicable, update minimum price and maximum price
69   } // repeat
70   write all counters onto database
71 }
```

B.1 DESCRIPTION OF THE ALGORITHM

The main routine is repeated 100 times (line 2) in order to provide sufficient data for statistical purposes. At the end of each repetition, the resulting trading data is saved onto the database (line 70), split between trading at thin spread (one tick) or wide spread (more than one tick) and category of aggressive and passive parties (LF for low-frequency trader, HF for high frequency trader), in order to apply statistical analysis later. The main routine runs over 10,000 cycles (line 9), each cycle representing a time period, whatever its length, during which either a limit or a market order is executed.

The first operation the main routine performs is randomly selecting the book it will act upon: either bid or ask (line 10). Then, if the bid-ask spread is larger than one tick (line 11), HF traders have less chances to be (randomly) selected than LF traders (line 12). The rationale for this choice is the following. The most well-known HFT feature is the tendency to quickly closing any open position, possibly with a, however small, profit. In order to do so, HF traders exploit the common swings shown by the markets but in order to take advantage of such minimal movements, the trading price must be very carefully chosen. If the bid-ask prices are, say, 100-101 and a HF trader buys at 101, chances are higher that a future upward movement will allow it to sell at 102 rather than at 103, necessary to make a profit if the initial spread was 100-102, instead. Conversely, if the spread is equal to one tick (line 15), then whether an HF or a LF traders is (randomly) decided by an algorithm that takes the QR parameter into account (line 16). For example, with QR=1 both type of traders have the same chances to be selected whereas with QR=10, a HF trader is ten times more likely to be selected than a low-frequency counterpart. Since HF traders regard wide spread too risky for their ultra-short-term strategy, when the spread is wide higher probability is given to (random) selection of limit order than to market orders (line 13). A simplifying assumption about liquidity has been made. All orders are supposed to only handle one lot of securities. A limit order will quote one lot at the given price and a market order will only trade one lot. The maximum liquidity allowed at any price level is 5 (although this is a parameter that can be changed). The order type is also selected using a random-number generation algorithm (line 17), yet assigning more chances to market order if the liquidity is in the upper half of the allowed range (i.e. if 3 through 5 limit orders are present at that price level). The order is then applied to the book selected earlier. However, in order to simulate quotes at a better price than top of book a special feature has been implemented. As described above, every limit order adds one lot to the book and every market order consumes exactly one lot; moreover each price level can only reach a maximum liquidity (set at initialisation time) and never trespassing it. So, when the top-of-book price level has reached the maximum liquidity, this is taken as sufficient consensus by the market on the soundness of that price to suggest the investor quoting the next price level (line 24).

APPENDIX C: PSEUDO-CODE FOR SIMULATION IN CHAPTER 6

The sequence of subroutine calls for the simulation of how High-Frequency Trading exploit arbitrage opportunities is as outlined by the following pseudo-code (comments and explanations within the pseudo-code are preceded by a double slash '//').

The Launch subroutine repeatedly calls a certain number of times the Arbitrage subroutine, which implements the simulation. The structure of the algorithm is as follows:

```
subroutine Launch{
    Read configuration parameters
    Requests # of repetitions //Default is 100 repetitions
    For each repetition {
        Call subroutine Arbitrage
        Copy computed average for the current repetition onto database
    } //End of For each repetition cycle
    Compute average and stdev of all repetitions
} //End of subroutine Launch

Subroutine Arbitrage
    For each iteration {
        Randomly select one security
        Randomly select price variation applied to the security (within +/- the maximum allowed value)
        Randomly select the market on which the price variation occurs (one or the other)
        Apply price variation to the selected security only in the selected market
        Select the trader fastest to arbitrage according to the HF-to-non-HF-trader ratio parameter
        Add the price variation and subtract transaction costs to the selected trader's cash account
    } //End of For each iteration cycle
    For each trader {
        Work out trader's account
    } //End of For each trader cycle
    Call ComputeAverage subroutine
```

```
}

subroutine ComputeAverage
    For each trader {
        Add up all HF trader's accounts
        Add up all LF trader's accounts
    } //End of For each trader cycle
    Divide total HF traders account value by number of HF traders
}
```

C.1 DESCRIPTION OF THE ALGORITHM

The simulation is made up of two markets trading the same set of securities. At each cycle a discrepancy arises in one of the markets by forcing the price change of one randomly-chosen security. All participants immediately notice that price difference but since the purpose of the simulation is to provide evidence of the principle, not working out the profit made through arbitrage, for sake of simplicity and without loss of generality, I set that one and only one trader is able to grasp the opportunity and that just one security is traded at every cycle. One randomly chosen trader buys the security in the market where the price is lower and sells it in the other market, so making a tiny profit and wiping out the arbitrage opportunity. At the end of the 10,000 cycles, the profits of the trade less the transaction costs are worked out for both the HF and LF trader communities. The simulation has been launched 100 times in order to provide a robust data set for statistical purposes.

APPENDIX D: DESCRIPTION OF THE ENVIRONMENT

D.1. INTRODUCTION

In the following paragraphs, I shall describe the functioning of the CME Globex electronic trading platform, its architecture, its rules and its main features. The source of most information contained in this section is the Chicago Mercantile Exchange Group website (www.cmegroup.com).

The CME Group exchanges offer a wide range of global benchmark products across all major asset classes, including futures and options based on interest rates, equity indexes, foreign exchange, energy, agricultural commodities, metals, weather and real estate. CME Group brings buyers and sellers together through its CME Globex electronic trading platform and its trading facilities in New York and Chicago. The CME Group also operates CME clearing, one of the largest central counterpart clearing services in the world, which provides clearing and settlement services for exchange-traded contracts, as well as for over-the-counter derivatives transactions through CME ClearPort.

The CME Globex was first introduced in operations in 1992 mostly on currency futures. The initial impact it had on the market was modestly successful but brighter days were about to come. The main breakthrough occurred in 1997, in the form of the E-mini Standard and Poor's (S&P) 500 Index futures contracts. By then, the S&P 500 Index futures was a well-established product since its introduction back in 1982, with the only slight drawback of being rather large in size: its value was a $500 multiplier of the S&P Index, reduced to $250 in 1997. The same year the E-mini version of that financial product was launched and it was offered exclusively on the CME Globex electronic trading platform, whence the 'E' prepended to the product name. The multiplier was reduced to $50, one tenth of the original S&P 500 futures contract, and its success was such that it is today one of the most liquid financial products in the world. The tick (the minimum amount the value of the E-mini S&P 500 Index futures contracts may fluctuate) is 0.25 index points, worth $12.50.

D.2. DESCRIPTION OF THE CME GLOBEX PLATFORM

The main source of information contained in this paragraph are Labuszewski et al. (2010) and Melamed (2009).

As of late 2009, 85% of all transactions going through the CME, did so via Globex. The primary location is Chicago but the CME offers access to it through a number of hubs located in Amsterdam, Dublin, London, Milan, New York, Paris, Sao Paulo, Seoul and Singapore. The equivalent of the old trading pit has now expanded to span from one side to the other of the entire world. The platform is continuously enhanced to serve customers with high-speed, high-volume capacity, improved options capabilities and a range of new products. CME Globex is an open access marketplace that allows customers to participate directly in the trading process, view the book of orders and prices for CME Group products and enter their own orders.

The most widely used order-matching algorithm across products is the 'First-In, First-Out', also known as FIFO queue. Orders are prioritised according to the price, that is, the highest bid and the lowest ask are placed first in the queue, with the others following in descending or ascending order (respectively). When two limit orders show the same price, then the FIFO algorithm applies. This is the algorithm used, among others, for the E-mini S&P 500 Index futures.

CME provides several order types, the most typical being limit orders, (pseudo) market orders, stop loss orders, market if touched orders, and one cancel the other orders. The Exchange offers an open access policy, whereby its customers can trade directly on the Globex platform, provided they have an account with a CME clearing member. Direct access is a source of extra revenue for exchanges offering it but it may come to the detriment of participants who do not enjoy the same access mode and have to relay their orders through brokers. In a world where micro- and even nano-seconds matter, this supplementary leap may cost dear in terms of execution rate. The platform also allows a special kind of orders, called market with protection, which allows investors to submit an order with a price buffer, known as protection points. If a bid (ask) order is entered at a value greater (less) than market price plus (minus) protection point, it is rejected by Globex. This permits protection against excessive volatility, like the one caused by fat finger issues. The width of the buffer depends on the product. It must be noticed that, as E-mini S&P 500 futures is one of the most liquid products, the protection applied to it is rather loose, compared to other products. Further to this, CME Globex generally enforces another form of protection, regardless of product, called price banding, with protection limits wider that those applicable to market with protection orders.

D.3. CME GLOBEX MAIN RULES

All CME Globex markets cycle through the daily order entry states, described below (The single-spaced sections are taken *verbatim* from the CME Reference Guide on www.cmegroup.com).

Pre-opening

A predetermined time before the trading session opens when customers can begin entering, modifying and cancelling orders for the next trading day, but no trades are executed.

Pre-opening/No-Cancel

A predetermined time before the session opens when customers can enter orders for the next trade date but cannot cancel or modify orders, and no trades are executed.

Open

The period of time when orders are sent and matched in real time, based on the product's trading times.

Pause

A pre determined time when customers can only cancel orders. No trades are executed.

Closed

This CME Globex state change, cancels, day orders and advance the trade date.

Post Close/Pre open (PCP)

This market state allows order placement, modification, and cancellation of GTC/GTD orders only. (Good 'Till Cancelled, or GTC, orders will remain in force until executed, cancelled or the contract expires. Good 'Till Date, or GTD, orders remain in force through the end of the specified date unless executed or cancelled, or the contract expires). No matching takes place and no action can be taken on non GTC/GTD orders.

Maintenance Period

Occurring between 16:15 Central Time (CT) and 16:45 (CT) Monday through Thursday.

Following are the main order types supported by the CME Globex platform (source: CME Reference Guide on www.cmegroup.com).

Limit order

A Limit order allows the buyer to define the maximum price to pay and the seller the minimum price to accept (the limit price).

Market with Protection order

Unlike a conventional Market order, where customers are at risk of having their orders filled at extreme prices, Market with Protection orders are filled within a predefined range of prices (the protected range). The protected range is typically the current best bid or offer, plus or minus 50 percent of the product's Non-Reviewable Trading Range.

No-Bust Range (Non-Reviewable Trading Range)

A range of prices used in determining if a potential error trade should be busted. The range is based on the true market price for the contract immediately before the error trade occurred, as determined by considering all relevant information, including the last trade price on the CME Globex platform, a better bid or offer price, a more recent price in a different contract month, the price of the same or a related contract established in open outcry trading and the prices of related contracts trading in other markets (e.g., cash FX and SGX Eurodollars). A trade may not be busted if it falls within the No Bust Range for that contract. No Bust Ranges vary by product.

Stop limit order

A resting Stop Limit order is triggered when the trigger price is traded in the market. The order then becomes a Limit order with the customer's specified limit price. The order is executed at all price levels between the trigger price and the limit price. A buy Stop Limit order must have a trigger price greater than the last traded price for the instrument. A sell Stop Limit order must have a trigger price lower than the last traded price for the instrument.

Stop with Protection order

A Stop with Protection is a Stop Limit order with the limit price calculated based on the trigger price, and the protected range. The protected range is typically the trigger price, plus or minus 50 percent of the Non-Reviewable Trading Range for that product. The limit price for a buy Stop with Protection will be calculated by adding the protected range to the trigger price. Likewise, the limit price for a sell Stop with Protection will be calculated by subtracting the protected range from the trigger price. Once the limit price for the order is calculated, it becomes a Stop with limit order in all respects.

D.4. FURTHER CME GLOBEX FEATURES

The CME Globex platform implements several features aiming to minimize excessive price movements and ensure fair markets (The single-spaced sections are taken *verbatim* from the CME Reference Guide on www.cmegroup.com).

Stop Spike logic

Stop Spike Logic prevents the excessive price movements caused by cascading stop orders by introducing a momentary pause in matching (Reserved State) when triggered stops would cause the market to trade outside predefined values (typically the same as the Non-Reviewable Trading Ranges). This momentary pause allows new orders to be entered and matched against the triggered stops in an algorithm similar to market opening. During the reserved period, customers can submit, modify and cancel all orders except Market Orders and Mass Quotes.

Velocity logic

Velocity Logic is designed to detect market movement of a predefined number of points either up or down within a predefined time. Velocity Logic introduces a momentary suspension in matching by transitioning the futures instrument(s) and related options into the Reserved/ Pause State. During the reserved period, customers can submit, modify and cancel all orders except Market Orders and Mass Quotes.

Price banding

To ensure fair and orderly markets, CME Group has a price banding mechanism in place that subjects all incoming electronic orders to price verification and rejects all orders with clearly erroneous prices. Price bands are monitored throughout the day by the CME Global Command Center (GCC) and adjusted if necessary.

Matching Algorithms

To ensure that customers get the best possible executions at the fairest prices, the CME Globex platform employs predefined sets of matching rules—algorithms— designed to best meet the needs of market participants in each product group.

FIFO

The FIFO algorithm uses price and time as the only criteria for filling an order. In this algorithm, all orders at the same price level are filled according to time priority; the first order at a price level is the first order matched.

As explained above, the CME Globex does not allow pure market orders. Therefore, if an investor wishes to hit a limit order it can only do so by entering a limit buy order at the prevailing ask or a limit sell order at the prevailing bid price. The result is a market order with the extra protection of being sure that the execution price shall not be worse than the one specified. This is a sound feature which prevents stub quotes being hit at absurdly high or low prices. During the Flash Crash Accenture saw its stock plunged, and hit, to as low as $0.01 whereas Sotheby's was bought at $99,999.99. Neither could have happened at CME because of the absence of pure market orders (and because of the other built-in protections). If order execution is critical, the investor can specify one or more ticks below (above) the prevailing bid (ask) price, increasing the probability of hitting the prevailing price even in case entire levels of liquidity had been consumed during the time the order reaches the exchange trading server.

This is exactly what CME Globex does in case of stop loss orders (called Stop Limit or Stop with Protection orders). A stop loss order is an order with two prices specified. One is the quoted price and the other is the price at which the order is turned to the opposite type if the market moves adversely. For example, a buy stop loss order is entered at 100 with stop loss at 98. If in the ask book there is an order at 100, the incoming order is executed and the stop loss order is activated. Then, should the market price fall to, or below, 98 the stop loss order will be immediately converted into an ask order at that price, to close the position with a still tolerable loss. Stop loss orders are not executed until a transaction occurs at the stop price. Since CME Globex does not allow market orders, the stop loss order is converted, as explained above, into a limit order with protection. However, since a stop loss order is entered to prevent excessive losses, the investor will be wise to set a reasonably wide protection point to allow for tolerable swings, otherwise the order could be so tightly protected that the stop loss executes too soon and unnecessarily causing (small) losses.

The idea behind establishing a protection point to stop loss orders is preventing a series of cascading stops. The triggering of a stop loss order could involve a quantity of securities that consumes all the liquidity at several price levels. If another stop loss order is awaiting at the new price hit, it will be triggered on its turn, potentially consuming liquidity at several price levels, and so on. With the protection implemented by CME Globex, the occurrence described is less likely, although not completely prevented. If stop loss levels are sufficiently close to each other, the protection mechanism may be unable to avoid the cascading stop loss phenomenon. The Stop Limit order feature might have played a role in the sharp price plunge experienced on May 6, 2010, and this was the subject of chapter 7. Another peculiar feature of CME Globex which played an important role on that day is the Stop Logic mechanism. The CME rules provide a Stop Logic under two different mechanisms.

Stop Spike Logic

Stop Spike Logic prevents the excessive price movements caused by cascading stop orders by introducing a momentary pause in matching (Reserved State) when triggered stops would cause the market to trade outside predefined values (typically the same as the Non-Reviewable Trading Ranges). This momentary pause allows new orders to be entered and matched against the triggered stops in an algorithm similar to market opening.

Velocity logic

Velocity Logic is designed to detect market movement of a predefined number of points either up or down within a predefined time. Velocity Logic introduces a momentary suspension in matching by transitioning the futures instrument(s) and related options into the Reserved/Pause State.

The predefined values for CME Globex to activate the Stop Spike Logic for the E-mini S&P 500 futures contracts was 6 index points in either direction. The Reserved State halts all trading for the financial instruments it applies to. During the halt period, orders are still booked as usual but they are not executed until the halt period has expired. This way, in the hope of the CME Globex designers, market participants shall have the time to consider the excessive swings and act accordingly. This may sounds too optimistic as five or ten seconds is such a short time but it was exactly the latter mechanism that saved the exchange (and possibly the US market at large) from the folly of the Flash Crash. May 6, 2010 was the first time in that year the Stop Logic functionality was activated.

D.5. CME GLOBEX ARCHITECTURE

The CME Group does not disclose details about CME Globex architecture. The only available information can be found on their website - and it is quite general.

In order to understand the components of the CME Globex platform one has to concentrate on the central part of the picture, the one labelled as Trading Platform. It is made up of six modules (as depicted in figure D1). The platform components are:

i) Historical Data & e-Data archives. Real-time market data is disseminated to market participant applications. Order book quotes, cancellations and trade data up to 10-deep is archived for for reselling and, in case of need, for post-mortem checks.

ii) Market Data is the module responsible of collecting data from the Matching, Pricing & Market Integrity module, and for dispatching it over to the front end applications and to the Quote Vendor System, for data dissemination purposes, and to the Historical Data & e-Data module, for archiving it.

iii) Matching, Pricing & Market Integrity. This module matches incoming with resting orders according to priorities by implementing the appropriate algorithms, handles schedules, and ensures integrity.

iv) Order Entry is the module that receives orders from CME customers and writes them onto the bid and ask books. It handles the interface with proprietary trading systems.

v) Drop Copy creates a hot-copy of the Order Entry data for failure & recovery purposes.

vi) The Credit Control & Risk Management module has several purposes, including trading customers and clearing firms protection.

Although the above only provides a virtual architecture with no intention to describe the hardware or software components of the platform, it is worth noticing the virtual, as well as physical, separation between the Order Entry (OE) and the Matching & Pricing (MP) modules. This has all important practical consequences. In case a new limit order arrives during execution of a pre-existing order at the same, or even worse, price level, since the Order Entry and the Matching & Pricing modules are mutually independent, the order in execution shall be completed (by the MP module) before the incoming order, serviced by the OE module, is read by the MP module. For example, the OE module writes to the bid book two limit orders at the highest price of, say, 100 for 20 stocks and 30 stocks, respectively. A sell market order for 50 stocks is posted 'at best' by the OE and it is matched and then executed against the bid limit orders at 100 by the MP. The execution of the market order is composed of two phases: the first matching the market order against the first 20 stocks and the second matching it against the 30 stocks. If another limit order at 101, that is, better than the prevailing price, arrives before the execution of the second phase, the MP module ignores this order until it has terminated processing the order it is being executed (at 100), even if the newly arrived price would provide a better match for the market order currently under execution.

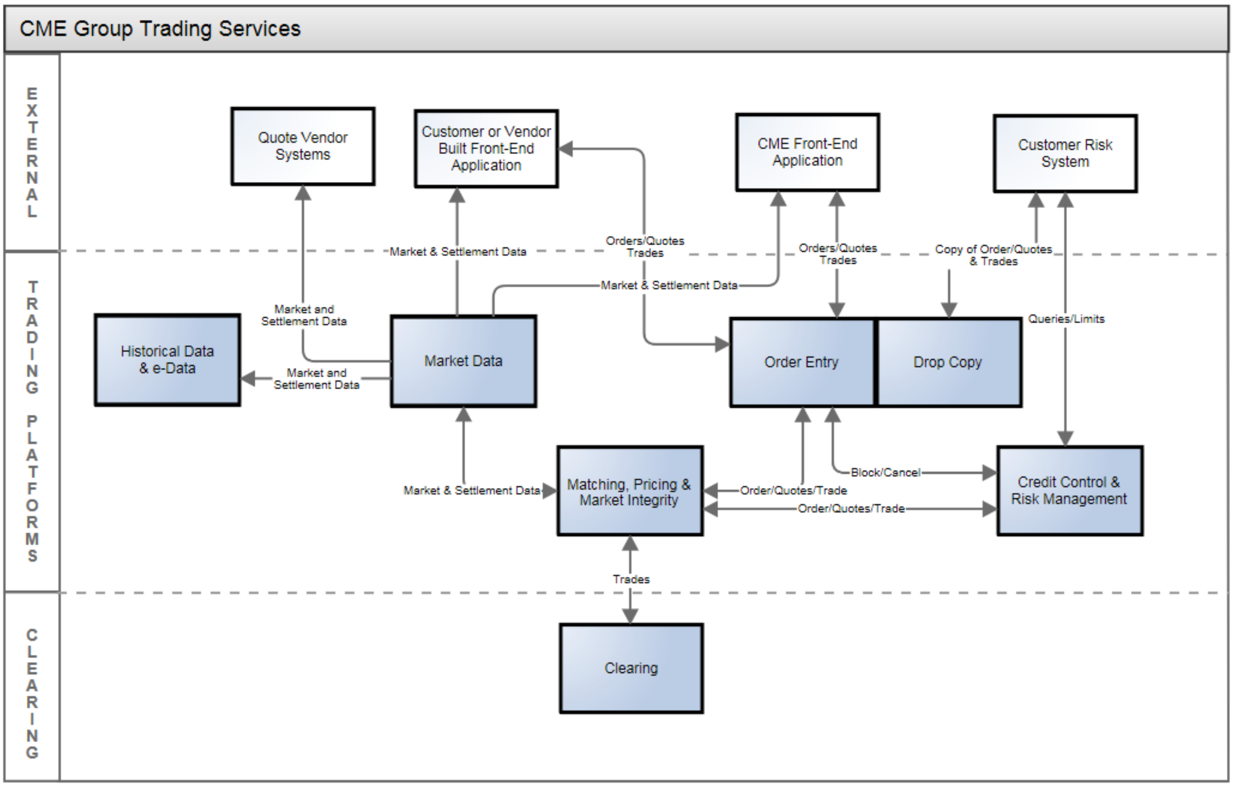


Figure D1. CME Globex high-level architecture (source: CME Group website)

REFERENCES

Abrol, Chesir and Mehta (2016): Abrol S., Chesir B., Mehta N. (2016). 'High Frequency Trading and US Stock Market Microstructure: A Study of Interactions between Complexities, Risks and Strategies Residing in U.S. Equity Market Microstructure'. *Financial Markets, Institutions & Instruments* 25(2) May. pp.107-165

Advent (2013): Advent Software Inc. (2013). *High-Frequency Trading: Useful Liquidity Tool or Weapon of Financial Mass Destruction*?. San Francisco: Advent Software Inc.

Ahlstedt and Villyson (2012): Ahlstedt J., Villysson J. (2012). *High frequency Trading*. Gothernburg: Chalmers University of Technology.

Aitken et al. (2012): Aitken M., de B. Harris F. H., McInish T., Aspris A., Foley S. (2012). 'High frequency trading - assessing the impact on market efficiency and integrity' *Foresight Driver Review DR28*. UK Government Office for Science

Aitken, Cumming and Zhan (2015): Aitken M., Cumming D., Zhan F. (2015). 'High frequency trading and end-of-day price dislocation'. *Journal of Banking & Finance* 59. pp.330–349. doi:10.1016/j.jbankfin.2015.06.011

Akram, Rime and Sarno (2009): Akram F., Rime D., Sarno L. (2009). 'Arbitrage in the foreign exchange market: Turning on the microscope'. *Journal of International Economics* 76. pp.237–253

Alawode and Al Sadek (2008): Alawode A., Al Sadek M. (2008). *What is Financial Stability?* Financial Stability Paper Series, no. 1. Bahrein: Central Bank of Bahrain

Aldridge (2010): Aldridge I. (2010). *High Frequency Trading*. Hoboken: John Wiley and Sons Inc.

Aldridge (2014): Aldridge I. (2014). 'High-Frequency Runs and Flash-Crash Predictability'. *The Journal of Portfolio Management*. 40(3) Spring. pp 113-123. doi:10.3905/jpm.2014.40.3.113

Aldridge and Krawciw [2015]: Aldridge I., Krawciw S. (2015). 'Aggressive High-Frequency Trading in Equities'. *Huffington Post Business*. Available at www.huffingtonpost.com/irene-aldridge/aggressive-highfrequency-_1_b_6698982.html?. Accessed on 05/02/2016

Andersen and Bondarenko (2014a): Andersen T.G., Bondarenko, O. (2014). 'VPIN and the flash crash *Journal of Financial Markets* 17. pp 1–46. doi:10.1016/j.finmar.2013.05.005

Andersen and Bondarenko (2014b): Andersen T.G., Bondarenko O. (2014). 'Reflecting on the VPIN dispute'. *Journal of Financial Markets* 17. pp 53–64. doi: 10.1016/j.finmar.2013.05.005

Anderson et al. (2015): Anderson N., Webber L., Noss J., Beale D., Crowley-Reidy L. (2015). *The resilience of financial market liquidity.* Financial Stability paper no. 34. London: Bank of England

Anderson (2016): Anderson B. (2016). 'Stock price leads and lags before the golden age of high-frequency trading'. *Applied Economics Letters* 23(3). pp 212-216. doi:10.1080/13504851.2015.1066481

Angel (2014): Angel J. (2014). 'When Finance Meets Physics: The Impact of the Speed of Light on Financial Markets and their Regulation'. *The Financial Review* 49(2) May. pp 271-281. doi: 10.1111/fire.12035

Aquilina and Ysusi (2016): Aquilina M., Ysusi C. (2016). *Are high-frequency traders anticipating the order flow? Cross-venue evidence from the UK market.* Occasional Paper 16. London: Financial Conduct Authority

Arnoldi (2016): Arnoldi J. (2016). 'Computer Algorithms, Market Manipulation and the Institutionalization of High Frequency Trading'. *Theory, Culture & Society* 33(1). pp 29–52. doi:10.1177/0263276414566642

Arnuk and Saluzzi (2009): Arnuk S., SaluzziJ. (2009). *Latency Arbitrage: The Real Power Behind Predatory High Frequency Trading*. Chaltam: Themis Trading LLC

Barker and Pomeranets (2011): Barker W., Pomeranets A. (2011). 'The Growth of High-Frequency Trading: Implications for Financial Stability'. *Financial System Review* June

Baron, Brogaard and Kirilenko (2012): Baron M., Brogaard J., Kirilenko A. (2012). *The Trading Profits of High Frequency Traders*. Princeton: Princeton University

Barrales (2012): Barrales E.O. (2012). 'Lessons from the flash crash for the regulation of high-frequency traders'. *Fordham Journal of Corporate and Financial Law* 17(4). pp 1195-1262

Benos and Sagade (2016): Benos E., Sagade S. (2016). 'Price discovery and the cross-section of high-frequency trading'. *Journal of Financial Markets* 30 September. pp 54–77. doi:10.1016/j.finmar.2016.03.004

Berman (2010): Berman G. (2010). *Speech by SEC Staff: Market Participants and the May 6 Flash Crash; 11th Annual SIFMA Market Structure Conference*. New York: Securities and Exchange Commission. Available at www.sec.gov/news/speech/2010/spch101310geb.htm. Accessed on 14/11/2013

Beunza, Millo and Pardo-Guerra (2012): Beunza D., Millo Y., Pardo-Guerra JP. (2012). 'Structured interviews of computer-based traders'. *Foresight Driver Review IN1*. UK Government Office for Science

Bhupathi (2010): Bhupathi T. (2010). 'Technology's Latest Market Manipulator? High Frequency Trading: The Strategies, Tools, Risks, and Responses'. *North Carolina Journal of Law & Technology* 11(2) Spring. pp 377-400

Biais, Foucault and Moinas (2014): Biais B., Foucault T., Moinas S. (2014). 'Equilibrium Fast Trading'. *Journal of Financial Economics* 116(2) May. pp. 292-313. doi: 10.1016/j.jfineco.2015.03.004

Blocher et al. (2016): Blocher J., Cooper R., Seddon J., Van Vliet B. (2016). 'Phantom Liquidity and High-Frequency Quoting'. *The Journal of Trading* 11(3) Summer. pp. 6-15. doi:10.3905/jot.2016.11.3.006

Bollen and Whaley (2015): Bollen N., Whaley R. (2015). 'The Journal of Futures Markets' 35(5). pp. 426–454. doi:10.1002/fut.21666

Brenner, Subrahmanyam and Uno (1990): Brenner M., Subrahmanyam M.G., Uno J. (1990). 'Arbitrage Opportunities in the Japanese Stock and Futures Markets'. *Financial Analysts Journal* 46(2) Mar-Apr. pp. 14-24

Brewer, Cvitanic and Plott (2013): Brewer P., Cvitanic J., Plott C. (2013). 'Market Microstructure Design And Flash Crashes: A Simulation Approach'. *Journal of Applied Economics* 16(2) November. pp. 223-250. doi:10.1016/S1514-0326(13)60010-0

Brogaard (2010): Brogaard J. (2010). *High Frequency Trading and Its Impact on Market Quality*. Evanston: Northwestern University

Brogaard (2011): Brogaard J. (2011). 'High frequency trading, information, and profits'. *Foresight Driver Review DR10* UK Government Office for Science

Brogaard et al. (2014): Brogaard J., Hendershott T., Hunt S., Latza T., Pedace L., Ysusi C. (2014). 'High-frequency trading and the execution costs of institutional investors'. *The Financial Review* 49(2) May. pp. 345-369. doi:10.1111/fire.12039

Brogaard, Hendershott and Riordan (2014): Brogaard J., Hendershott T., Riordan R. (2014). 'High Frequency Trading and Price Discovery'. *The Review of Financial Studies* 27(8). pp. 2267-2306. doi:10.1093/rfs/hhu032

Brogaard, Moyaert and Riordan (2014): Brogaard J., Moyaert T., Riordan R. (2014). *High Frequency Trading and Market Stability*. Washington: University of Washington

Buchanan (2009): Buchanan M. (2009). 'Meltdown modelling'. *Nature* 460 August. pp. 680-682

Buchanan (2015): Buchanan M. (2015). 'Trading at the speed of light'. *Nature* 518 February. pp. 161-163

Budimir and Schweickert (2009): Budimir M., Schweickert U. (2009). 'Latency in Electronic Securities Trading: A Proposal for Systematic Measurement'. *The Journal of Trading* Summer. pp. 47-55

Budish, Cramton and Shim (2015): Budish E., Cramton P., Shim J. (2015). 'The High-Frequency Trading Arms Race: Frequent Batch Auctions As A Market Design Response'. *The Quarterly Journal of Economics* 130(4) November. pp. 1547–1621. doi:10.1093/qje/qjv027

Bullock (2011): Bullock S. (2011). 'Prospects for large-scale financial systems simulation'. *Foresight Driver Review DR14*. UK Government Office for Science

Carrion (2013): Carrion A. (2013). 'Very fast money: High-frequency trading on the NASDAQ'. *Journal of Financial Markets* 16. pp. 680–711. doi:10.1016/j.finmar.2013.06.005

Carter (1989): Carter C. (1989). 'Arbitrage opportunities between thin and liquid futures markets'. *Journal of Futures Markets* 9(4) August. pp. 347–353

Cartlidge and Cliff (2012): Cartlidge J., Cliff D. (2012). 'Exploring the 'robot phase transition' in experimental human-algorithmic markets'. *Foresight Driver Review DR25*. UK Government Office for Science

Castiglione (2006): Castiglione F. (2006). *Agent based modelling*. Scholarpedia 1(10):1562.

Available at: www.scholarpedia.org/article/Agent_based_modeling doi:10.4249/scholarpedia.1562. Accessed on 14/8/2015

CFTC-SEC (2010a): Commodity Futures Trading Commission, Securities and Exchange Commission. *Preliminary Findings Regarding the Market Events of May 6, 2010*. Washington: Commodity Futures Trading Commission, Securities and Exchange Commission

CFTC-SEC (2010b): Commodity Futures Trading Commission, Securities and Exchange Commission. *Findings Regarding the Market Events of May 6, 2010*. Washington: Commodity Futures Trading Commission, Securities and Exchange Commission

CFTC-SEC (2011): Commodity Futures Trading Commission, Securities and Exchange Commission. *Recommendations regarding Regulatory Responses to the Events of May 6, 2010*. Washington: Commodity Futures Trading Commission, Securities and Exchange Commission

Chaboud et al. (2014): Chaboud A., Chiquoine B., Hjalmarsson E., Vega C. (2013). 'Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market'. *The Journal of Finance* 69(5) October 2014, (2045–2084), doi:10.1111/jofi.12186

Cliff (2011a): Cliff D. (2011). 'Regulatory scrutiny of algorithmic trading systems: an assessment of the feasibility and potential economic impact'. *Foresight Driver Review EIA16*. UK Government Office for Science

Cliff (2011b): Cliff D. (2011). 'Market-making obligations and algorithmic trading systems A feasibility assessment of the March 2012 draft of MiFID2 Article 17(3)'. *Foresight Driver Review EIA19*. UK Government Office for Science

Cliff, Brown and Treleaven (2010): Cliff D., Brown D., Treleaven P. (2010). 'Technology Trends in the Financial Markets: A 2020 Vision'. *Foresight Driver Review DR3*. UK Government Office for Science

Cliff and Northrop (2010): Cliff D., Northrop L. (2010). 'The Global Financial Market: an Ultra-Large-Scale System perspective'. *Foresight Driver Review DR4*. UK Government Office for Science

CME (2010): CME Group (2010). *What Happened on May 6th?*. Chicago: Chicago Mercantile Exchange

Cohen and Szpruch (2012): Cohen S., Szpruch L. (2012). 'A limit order book model for latency arbitrage'. *Mathematics and Financial Economics* 6(3) June. pp. 211-227. doi:10.1007/s11579-012-0082-5

Colliard (2016): Colliard JE. (2016). 'Catching Falling Knives: Speculating on Liquidity Shocks'. *Management Science* (forthcoming). doi:10.1287/mnsc.2016.2440

Conrad, Wahal and Xiang (2015): Conrad J., Wahal S., Xiang J. (2015). 'High-frequency quoting, trading, and the efficiency of prices'. *Journal of Financial Economics* 116. pp. 271–291. doi:10.1016/j.jfineco.2015.02.008

Copeland and Galai (1983): Copeland T., Galai D. (1983) 'Information Effects on the Bid-Ask Spreads'. *Journal of Finance* 38. pp.1457–1469

Cvitanić and Kirilenko (2010): Cvitanić J., Kirilenko A. (2010). *High Frequency Traders and Asset Prices*. Pasadena: California Institute of Technology

Danielsson (2013): Danielsson J. (2013). *Global Financial Systems Stability and Risk*. Harlow: Pearson

Danielsson and Zer (2012): Danielsson J., Zer I. (2012). **'**Systemic risk arising from computer based trading and connections to the empirical literature on systemic risk'. *Foresight Driver Review DR29*. UK Government Office for Science

Darley and Outkin (2007): Darley V., Outkin A. (2007). *A NASDAQ Market Simulation Vol. I*. Singapore: World Scientific Publishing Co. Pte Ltd.

Davis, Van Ness and Van Ness (2014): Davis R., Van Ness B., Van Ness R. (2014). 'Clustering of Trade Prices by High-Frequency and Non–High-Frequency Trading Firms'. *The Financial Review* 49. pp. 421–433

De Luca et al. (2011): De Luca M., Szostek C., Cartlidge J., Cliff D. (2011). 'Studies of interactions between human traders and Algorithmic Trading Systems'. *Foresight Driver Review DR13*. UK Government Office for Science

Diaz-Rainey, Ibikunle and Mention (2015): Diaz-Rainey I., Ibikunle G., Mention, A. (2015). 'The technological transformation of capital markets'. *Technological Forecasting & Social Change* 99. pp. 277–284. doi:10.1016/j.techfore.2015.08.006

Ding, Hanna and Hendershott (2014): Ding S., Hanna J., Hendershott T. (2014). 'How Slow Is the NBBO? A Comparison with Direct Exchange Feeds'. *The Financial Review* (49). pp. 313–332

Durden (2010): Durden T. (2010). *How HFT Quote Stuffing Caused The Market Crash Of May 6. And Threatens To Destroy The Entire Market At Any Moment.* Available at www.zerohedge.com /article/how-hft-quote-stuffing-caused-market-crash-may-6-and-threatens-destroy-entire-market-any-mom. Accessed on 07/08/2013

Durden (2012): Durden T. (2012). *Was the SEC 'Explanation' Of the Flash Crash Maliciously Fabricated or Completely Flawed Out Of Plain Incompetence?*. Available at www.zerohedge.com /news/was-sec-explanation-flash-crash-maliciously-fabricated-or-completely-flawed-due-plain-incompete. Accessed on 14/08/2013

Durden (2013): Durden T. (2013). *Flash Crash Mystery Solved*. Available at www.zerohedge.com /news/2013-03-27/flash-crash-mystery-solved. Accessed on 14/08/2013

Durden (2014): Durden T. (2014). *High Frequency Trading: All You Need To Know*. Available at www.zerohedge.com /news/2014-04-06/high-frequency-trading-all-you-need-know. Accessed on 08/04/2015

Durenard (2013): Durenard E. (2013). *Professional Automated Trading*. Hoboken: John Wiley and Sons Inc.

Easley, Lopez de Prado and O'Hara (2011): Easley D., Lopez de Prado M., O'Hara M. (2011). 'The Microstructure of the 'Flash Crash': Flow Toxicity, Liquidity Crashes, and the Probability of Informed Trading'. *The Journal of Portfolio Management* 37(2) Winter. pp. 118-128.

doi:10.3905/jpm.2011.37.2.118

Easley, Lopez de Prado and O'Hara (2012): Easley D., Lopez de Prado M., O'Hara M. (2012). 'Flow Toxicity and Liquidity in a High-frequency World'. *The Review of Financial Studies* 25(5) May. pp. 1457-1493

Easley, Lopez de Prado and O'Hara (2013): Easley D., Lopez de Prado M., O'Hara M. (2013). 'The Volume Clock: Insights into the High-Frequency Paradigm'. *The Journal of Portfolio Management*. 39(1) Fall. pp. 19-29. doi:10.3905/jpm.2012.39.1.019

Easley, Lopez de Prado and O'Hara (2014): Easley D., Lopez de Prado M., O'Hara M. (2014). 'VPIN and the Flash Crash: A rejoinder'. *Journal of Financial Markets* (17). pp. 47–52. doi:10.1016/j.finmar.2013.05.005

Easley, Lopez de Prado and O'Hara (2016): Easley D., Lopez de Prado M., O'Hara M. (2016). 'Differential Access to Price Information in Financial Markets'. *Journal of Financial and Quantitative Analysis* Forthcoming

Essendorfer, Diaz-Rainey, Falta (2015): Essendorfer S., Diaz-Rainey I., Falta M. (2015). 'Creative destruction in Wall Street's technological arms race: Evidence from patent data'. *Technological Forecasting and Social Change* 99. pp. 300-316). doi:10.1016/j.techfore.2014.11.012

Fama (1965): Fama E. (1965). 'The Behavior Of Stock Market Prices'. *The Journal of Business* 38(1) January. pp. 34-105

Fama (1970): Fama E. (1970). 'Efficient Capital Markets: A Review of Theory and Empirical Work'. *The Journal of Finance* 25(2) May. pp. 383-417

Farmer and Skouras (2012a): Farmer J.D., Skouras S. (2012). 'Minimum resting times and transaction-to-order ratios: review of Amendment 2.3.f and Question 20'. *Foresight Driver Review EIA2*. UK Government Office for Science

Farmer and Skouras (2012b): Farmer J.D., Skouras S. (2012). 'Review of the benefits of a continuous market vs. randomised stop auctions and of alternative Priority Rules (policy options 7 and 12)'. *Foresight Driver Review EIA11*. UK Government Office for Science

Farmer and Skouras (2013): Farmer J.D., Skouras S. (2013). 'An Ecological Perspective on The Future of Computer Trading'. *Quantitative Finance* 13(3). pp. 325-346. doi:10.1080/14697688.2012.757636

Fell and Schinasi (2005): Fell J., Schinasi G. (2005). 'Assessing Financial Stability: Exploring the Boundaries of Analysis'. *National Institute Economic Review* 192(1) April. pp. 102-117

Foot (2003): Foot M. (2003). *What is financial stability and how we get it?* The Roy Bridge Memorial Lecture. London: The Financial Service Authority. Available at: http://www.fsa.gov.uk/library/communication/speeches/2003/sp122.shtml. Accessed on 20/4/2016

Foresight (2012): Foresight (2012). 'The Future of Computer Trading in Financial Markets, Final Project Report'. UK Government Office for Science, London

Foucault (2012): Foucault T. (2012). 'Pricing Liquidity in Electronic Markets'. *Foresight Driver Review DR18*. UK Government Office for Science

Foucault, Kadan and Kendel (2013): Foucault T., Kadan O., Kandel E. (2013). 'Liquidity Cycles and Make/Take Fees in Electronic Markets'. *The Journal of Finance* 68(1) February. pp. (299-341). doi: 10.1111/j.1540-6261.2012.01801.x

Foucault, Hombert and Roşu (2016): Foucault T., Hombert J., Roşu I. (2016). 'News Trading and Speed'. *The Journal of Finance* 71(1) February. pp. 335-382 doi:10.1111/jofi.12302

Friederich and Payne (2011): Friederich S., Payne R. (2011). 'Computer based trading, liquidity and trading costs'. *Foresight Driver Review DR5*. UK Government Office for Science

Friederich and Payne (2012): Friederich S., Payne R. (2012). 'Computer-based trading and market abuse'. *Foresight Driver Review DR20*. UK Government Office for Science

Friederich and Payne (2015): Fiederich S., Payne R. (2015). 'Order-to-trade ratios and market liquidity'. *Journal of Banking & Finance* 50 pp. 214–223. doi:10.1016/j.jbankfin.2014.10.005

Frino and Lepone (2012): Frino A., Lepone A. (2012). 'The impact of high frequency trading on market integrity: an empirical examination'. *Foresight Driver Review DR24*. UK Government Office for Science

Gagnon and Karolyi (2010): Gagnon L., Karolyi A. (2010). 'Multi-market trading and arbitrage'. *Journal of Financial Economics* 97(1). pp. 53-80

Garman (1976): Garman M. (1976). 'Market microstructure'. *Journal of Financial Economics* 3(3). pp. 257–275

Garvey and Murphy (2006): Garvey R., Murphy A. (2006). 'Crossed Markets: Arbitrage Opportunities in Nasdaq Stocks'. *The Journal of Alternative Investments* 9(2) Fall. pp. 46–58. doi: 10.3905/jai.2006.655936

Ghadhab and Hellara (2015): Ghadhab I., Hellara S. (2015). 'The law of one price, arbitrage opportunities and price convergence: Evidence from cross-listed stocks'. *Journal of Multinational Financial Management* (31). pp. 126–145. doi: 0.1016/j.mulfin.2015.05.002

Giffords (2010): Giffords B. (2010). 'What Just Happened?'. *Automated Trader Magazine* (18) Q3

Gilbert (2008): Gilbert N. (2008). *Agent-based models; Quantitative Applications in Social Sciences* 153. Los Angeles: Sage Publications Inc.

Gleick (2008): Gleick, J. (2008). *Chaos – Making a New Science*. New York: Penguin Books

Goldstein, Kumar and Graves (2014): Goldstein M., Kumar P., Graves F. (2014). 'Computerized and High-Frequency Trading'. *The Financial Review (*49). pp. 177–202

Golub, Keane and Poon (2012): Golub A., Keane J., Poon S. (2012). *High Frequency Trading and Mini Flash Crashes*. Manchester: Unversity of Manchester. Available at SSRN: ssrn.com/abstract=2182097, accessed on 28/4/2014. doi: 10.2139/ssrn.2182097

Gomber et al. (2011): Gomber P., Arndt B., Lutat M., Uhle T. (2011). High-Frequency Trading. Frankfurt am Main: Goethe Universität

Groth (2011): Groth S. (2011). *Does Algorithmic Trading Increase Volatility? Empirical Evidence* *from the Fully-Electronic Trading Platform Xetra* Wirtschaftsinformatik Proceedings Paper 112. Frankfurt am Main: Goethe Universität

Grossman and Stiglitz (1980): Grossman S., Stiglitz J. (1980). 'On the Impossibility of Informationally Efficient Markets'. *The American Economic Review* 70(3) June. pp. 393-408

Gsell (2008): Gsell M. (2008). *Assessing the Impact of Algorithmic Trading on Markets: A Simulation Approach*. Frankfurt am Main: Goethe Universität

Gyurkó (2010): Gyurkó L. (2010). 'The evolution of algorithmic classes'. *Foresight Driver Review DR17*. UK Government Office for Science

Hagströmer and Nordén (2013): Hagströmer B., Nordén L. (2013). 'The diversity of high frequency traders'. *Journal of Financial Markets* 16(4) November. pp. 741-770

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

Haldane (2011): Haldane A. (2011). *The race to zero* Speech held at the International Economic Association Sixteenth World Congress. Beijing: Bank of England

Hanif (2012): Hanif A. (2012). *Colocation and Latency Optimization* RN/12/04. London: University College London

Hanson (2016): Hanson T. (2016). 'High frequency traders in a simulated market'. *Review of Accounting and Finance* 15(3). pp. 329-351. doi: 10.1108/RAF-02-2015-0023

Harris (2013): Harris L. (2013). 'What To Do About High-Frequency Trading'. *Financial Analyst Journal* 69(2) March/April. pp. 6-9. doi: 10.2469/faj.v69.n2.6

Hasbrouck and Saar (2009): Hasbrouck J., Saar G. (2009). 'Technology and liquidity provision: the blurring of traditional definitions'. *Journal of Financial Markets* (12). pp. 143-172

Hasbrouck and Saar (2013): Hasbrouck J., Saar G. (2013). 'Low-Latency Trading'. *Journal of Financial Markets* 16(4) November. pp. 646-679

Hendershott and Moulton (2011): Hendershott T., Mouton P.C. (2011). 'Automation, speed, and stock market quality: The NYSE's Hybrid'. *Journal of Financial Markets* (14). pp. 568-604

Hendershott and Riordan (2013): Hendershott T., Riordan R. (2013). 'Algorithmic Trading and the Market for Liquidity'. *Journal of Financial and Quantitative Analysis* 48(4) August. pp. 1001–1024. doi:10.1017/S0022109013000471

Hendershott, Jones and Menkveld (2011): Hendershott T., Jones C., Menkveld A. (2011). ‘Does Algorithmic Trading Improve Liquidity?’. *The Journal of Finance* 66(1) February 2011. pp. 1-33. doi:10.1111/j.1540-6261.2010.01624.x

Hendershott (2011): Hendershott T. (2011). 'High frequency trading and price efficiency'. *Foresight Driver Review DR12*. UK Government Office for Science

Hens, Herings and Predtetchinskii (2006): Hens T., Herings J.J, Predtetchinskii A. (2006). 'Limits to Arbitrage When Market Participation Is Restricted'. *Journal of Mathematical Economics* 42(4-5).

pp. 556-564

Hoffmann (2014): Hoffmann P. (2014). 'A dynamic limit order market with fast and slow traders'. *Journal of Financial Economics* 113(1) July. pp. 156–169. doi:10.1016/j.jfineco.2014.04.002

Hruska and Linnertova (2015): Hruska J., Linnertova D. (2015). 'Liquidity of the European stock markets under the influence of HFT'. *Procedia Economics and Finance* (26). pp. 375–381. doi:10.1016/S2212-5671(15)00867-9

Jain (2005): Jain P. (2005). 'Financial Market Design and the Equity Premium: Electronic versus Floor Trading'. *The Journal of Finance* 60(6) December

Jarnecic and Snape (2014): Jarnecic E., Snape M. (2014). 'The Provision of Liquidity by High-Frequency Participants'. *The Financial Review* (49) pp. 371–394

Jarrow and Protter (2012): Jarrow R., Protter P. (2012). 'A Dysfunctional Role of High Frequency Trading in Electronic Markets'. *International Journal of Theoretical and Applied Finance* 15(3). doi:0.1142/S021902491250022

Johansen and Sornette (2010): Johansen A., Sornette D. (2010). 'Shocks, Crashes and Bubbles in Financial Markets'. *Brussels Economic Review* 53(2) Summer. pp. 201-253

Johnson and Tivnan (2012): Johnson N., Tivnan B. (2012). 'Mechanistic origin of dragon-kings in a population of competing agents**'.** *European Physics Journal - Special Topics* 205(65). pp. 65-78)

Johnson and Zhao (2012): Johnson N., Zhao G. (2012). 'Brave new world: quantifying the new instabilities and risks arising in subsecond algorithmic trading'. *Foresight Driver Review DR27*. UK Government Office for Science

Johnson et al. (2013): Johnson N., Zhao G., Hunsader E., Meng J., Ravinder A., Carran S., Tivnan B. (2013). 'Abrupt rise of new machine ecology beyond human response time'. *Scientific Reports* 3(2627). doi:10.1038/srep02627

Jovanovic and Menkveld (2016): Jovanovic B., Menkveld A. (2016). *Middlemen in Limit-Order Markets*. Available at: ssrn.com/abstract=1624329. Accessed on 18/12/2014. doi:10.2139/ssrn.1624329

Jung Lee (2015): Jung Lee E. (2015). 'High Frequency Trading in the Korean Index Futures Market'. *The Journal of Futures Markets* 35(1). pp. 31–51. doi:10.1002/fut.21640

Kalejian and Mukerji (2016): Kelejian H.H., Mukerji P. (2016). 'Does high frequency algorithmic trading matter for non-AT investors?'. *Research in International Business and Finance* (37). pp. 78–92. doi:10.1016/j.ribaf.2015.10.014

Kato, Limm and Schallheim (1990): Kato K., Linn S., Schallheim J. (1990). 'Are there arbitrage opportunities in the market for American depository receipts?'. *Journal of International Financial Markets, Institutions & Money* 1(1). pp. 73-89

Keller (2003): Keller E. (2003). 'Models, simulation, and computer experiments'. In Radder H. (ed.) *The philosophy of scientific experimentation*. Pittsburgh: University of Pittsburgh Press. pp. 198-215

Kirkpatrick and Dahlquist (2007): Kirkpatrick C., Dahlquist J. (2007). *Technical Analysis - The Complete Resource For Financial Market Technicians*. Upper Saddle River: FT Press

Kirilenko and Lo (2013): Kirilenko A., Lo A. (2013). 'Moore's Law versus Murphy's Law: Algorithmic Trading and Its Discontents'. *Journal of Economic Perspectives* 27(2) Spring. pp. 51-72. 10.1257/jep.27.2.51

Kirilenko et al. (2011): Kirilenko A., Kyle A., Samadi M., Tuzun T. (2011). *The Flash Crash: The Impact of High Frequency Trading on an Electronic Market*; Washington: Commodity Futures Trading Commission. Available at SSRN: ssrn.com/abstract=1686004, accessed on 14/11/2013

Kolb, Gay and Jordan (1982): Kolb R., Gay G., Jordan J. (1982). 'Are there arbitrage opportunities in the treasury-bond futures market?'. *Journal of Futures Markets* 2(3) Autumn. pp. 217–229

Krishnamurti (2009): Krishnamurti C. (2009). 'Introduction to Market Microstructure'. In Vishwanath S.R., Krishnamurti C. (eds.) *Investment Management: A Modern Guide to Security Analysis and Stock Selection*. Berlin Heidelberg: Springer-Verlag. pp. 13-29. doi:10.1007/978-3-540-88802-4_2

Labuszewski et al. (2010): Labszewski J., Nyhoff J., Co R., Peterson P. (2010). *The CME Group Risk Management Handbook*. Hoboken: John Wiley and Sons Inc.

Laughlin, Aguirre and Grundfest (2014): Laughlin G., Aguirre A., Grundfest J. (2014). 'Information Transmission between Financial Markets in Chicago and New York'. *The Financial Review* 49. pp. 283–312

Leland (2011): Leland H. (2011). 'Leverage, Forced Asset Sales and Market Stability: Lessons from Past Market Crises and the Flash Crash'. *Foresight Driver Review DR9*. UK Government Office for Science

Leshik (2011): Leshik E., Cralle J. (2011). *An Introduction to Algorithmic Trading*. Chichester: John Wiley and Sons Ltd.

Lehtinen and Kuorikoski (2007): Lehtinen A., Kuorikoski J. (2007). 'Computing the perfect model: Why do economists shun simulation?'. *Philosophy of Science* (74). pp. 304-329

Levens (2015): Levens T. (2015). 'Too Fast, Too Frequent? High-Frequency Trading and Securities Class Actions'. *The University of Chicago Law Review* 82(3) Summer. pp. 1511-1557

Lewis (2014a): Lewis M. (2014). *'Flash Boys': A Wall Street Revolt.* New York: W. W. Norton & Co.

Lewis (2014b): Lewis M. (2014). *An Adaptation From 'Flash Boys: A Wall Street Revolt'*. Available at: www.nytimes.com/2014/04/06/magazine/flash-boys-michael-lewis.html. Accessed on 3/2/2015

Linton and O’Hara (2012): Linton O., O'Hara M. (2012). 'The impact of computer trading on liquidity, price efficiency/discovery and transaction costs'. *Foresight Driver Review WP2*. UK Government Office for Science

MacKenzie (2015): MacKenzie D. (2015). 'Mechanizing the Merc: The Chicago Mercantile Exchange and the Rise of High- Frequency Trading'. *Technology and Culture* 56(3) July. pp. 646-

675. doi:10.1353/tech.2015.0102

Madhavan (2012): Madhavan A. (2012). 'Exchange Traded Funds, Market Structure and the Flash Crash'. *Financial Analysts Journal* 68(4). pp. 20–35

Malinova and Park (2015): Malinova K., Park A. (2015). 'Subsidizing Liquidity: The Impact of Make/Take Fees on Market Quality'. *The Journal of Finance* 70(2) April. pp. 509-536. doi: 10.1111/jofi.12230

Malkiel (2003): Malkiel B. (2003). 'The Efficient Market Hypothesis and Its Critics'. *Journal of Economic Perspectives* 17(1) Winter. pp. 59-82

Manahov and Hudson (2014): Manahov V., Hudson R. (2014). 'The implications of high frequency trading on market efficiency and price discovery'. *Applied Economics Letters* 21(16). pp. 1148-1151. doi:10.1080/13504851.2014.914135

Manahov, Hudson and Gebka (2014): Manahov V., Hudson R., Gebka B. (2014). 'Does high frequency trading affect technical analysis and market efficiency? And if so, how?'. *Journal of International Financial Markets, Institutions & Money* 28. pp. 131-157

Markham (2015): Markham J. (2015). 'High Speed Trading on Stock and Commodity Markets—From Courier Pigeons to Computers'. *San Diego Law Review* 52(3) September. pp. 555-618

McGowan (2010): McGowan M.J. (2010). 'The Rise of Computerized High Frequency Trading: Use and Controversy'. *Duke Law and Technology Review* 16

McInish, Upson and Wood (2012): McInish T., Upson J., Wood R. (2012). *The Flash Crash: Trading Aggressiveness, Liquidity Supply, and the Impact of Intermarket Sweep Orders*. El Paso: University of Texas. Available at ssrn.com/abstract=1629402, accessed on 06/03/2016

McInish and Upson (2012): McInish T., Upson J. (2012). *Strategic Liquidity Supply in a Market with Fast and Slow Traders*. El Paso: University of Texas

McInish and Upson (2013): McInish T., Upson J. (2013). 'The Quote Exception Rule: Giving High Frequency Traders An Unintended Advantage'. *Financial Management* 42(3). pp. 481-501. doi:10.1111/fima.12017

Melamed (2009): Melamed L. (2009). *For Crying Out Loud: From Open Outcry to the Electronic Screen*. Hoboken: John Wiley & Sons Inc.

Menkveld (2014): Menkveld A. (2014). 'Electronic trading and market structure'. *The Financial Review* 49(2) May. pp. 333–344. doi:10.1111/fire.12038

Menkveld (2013): Menkveld A. (2013). 'High Frequency Trading and the New-Market Makers'. *Journal of Financial Markets* 16(4) November. pp. 712-740

Menkveld and Yueshen (2016): Menkveld A., Yueshen B.Z. (2016). *The Flash Crash: A Cautionary Tale about Highly Fragmented Markets*. Available at: ssrn.com/abstract=2243520 Accessed on 7/5/2016. doi: 10.2139/ssrn.2243520

Menkveld and Zoican (2016): Menkveld A., Zoican M. (2016). *Need for Speed? Exchange Latency and Liquidity* Tinbergen Institute Discussion Paper 14-097/IV/DSF78. Available at:

ssrn.com/abstract=2442690. Accessed on 7/5/2016. doi:10.2139/ssrn.2442690

Mitra et al. (2016): Mitra G., diBartolomeo D., Banerjee A., Yu X. (2016). 'Automated Analysis of News to Compute Market Sentiment: Its Impact on Liquidity and Trading'. In Mitra, G. And Yu, X. (eds.) *Handbook of Sentiment Analysis in Finance*. published by the editors. Ch. 22

Moallemi and Sağlam (2013): Moallemi C., Sağlam M. (2013). *The Cost of Latency in High-Frequency Trading*. Available at ssrn.com/abstract=1571935. Accessed on 28/4/2014

Murata (1989): Murata T. (1989). 'Petri Nets: Properties, Analysis and Applications'. *Proceedings of the IEEE* 77(4) April

Myers and Gerig (2014): Myers B., Gerig A. (2014). 'Simulating the Synchronizing Behavior of High-Frequency Trading in Multiple Markets'. In Bera A., Ivliev S., Lillo F. *Financial Econometrics and Empirical Market Microstructure*. Berlin: Springer. pp. 207-213. doi:10.1007/978-3-319-09946-0_13

Nanex (2010a): Nanex (2010). *Analysis of the 'Flash Crash' - Intro Empirical Analysis*. Available at: www.nanex.net /20100506/FlashCrashAnalysis_Intro.html. Accessed on 23/05/2013

Nanex (2010b): Nanex (2010). *Analysis of the 'Flash Crash' - Part 1 Empirical Analysis*. Available at: www.nanex.net /20100506/FlashCrashAnalysis_CompleteText.html. Accessed on 07/08/2013

Nanex (2010c): nanex (2010). *Analysis of the 'Flash Crash' - Part 4 Quote Stuffing*. Available at: www.nanex.net /20100506/FlashCrashAnalysis_Part4-1.html. Accessed on 14/11/2013

Newbold (1995): Newbold P. (1995). *Statistics for Business and Economics*. Englewood Cliffs: Prentice-Hall International Inc.

O'Hara (1995): O'Hara M. (1995). *Market Microstructure Theory*; Malden: Blackwell Publishing Ltd.

O'Hara (2015): O'Hara M. (2015). 'High frequency market microstructure'. *Journal of Financial Economics* 116. pp. 257-270. doi:10.1016/j.jfineco.2015.01.003

Paddrick et al. (2012): Paddrik M., Hayes R. Jr., Todd A., Yang S., Beling P., Scherer W. (2012). *An agent based model of the E-Mini S&P 500 applied to flash crash analysis. Proceedings of IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. IEEE 29-30/3/2012. New York, IEEE.

Pandey and Wu (2015): Pandey V., Wu C. (2015). 'Investors May Take Heart: A Game Theoretic View of High Frequency Trading'. *The Journal of Financial Planning* May. pp. 53-57

Pirrong (2014): Pirrong C. (2014). 'Pick Your Poison—Fragmentation or Market Power? An Analysis of RegNMS, High Frequency Trading, and Securities Market Structure'. *Journal of Applied Corporate Finance* 26(2) Spring. pp. 8-14. doi: 10.1111/jacf.12061

Popova-Zeugmann (2013): Popova-Zeugmann L. (2013). *Time and Petri Nets*. Berlin Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-41115-1

RegNMS (2005): Securities and Exchange Commission (2005). *Regulation National Market System*. Washington: Securities and Exchange Commission

Reisig (2013): Reisig W. (2013). *Understanding Petri Nets-Modeling Techniques, Analysis Methods, Case Studies*. Berlin Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-33278-4

Reiss (2011): Reiss J. (2011). 'A Plea for (Good) Simulations: Nudging Economics Toward an Experimental Science'. *Simulation & Gaming* 42(2). pp. 243-264. doi:10.1177/1046878110393941

Samuelson (1965): Samuelson P. (1965). 'Proof That Properly Anticipated Prices Fluctuate Randomly'. *Industrial Management Review* 6(2).Spring. pp. 41-49

Sandas (2001): Sandas P. (2001). 'Adverse Selection and Competitive Market Making: Empirical Evidence from a Limit Order Market'. *Review of Financial Studies* 14(3). pp. 705-734. doi:10.1093/rfs/14.3.705

Schapiro (2010): Schapiro M. (2010). *Testimony Concerning the Severe Market Disruption on May 6, 2010 Before the Subcommittee on Capital Markets Insurance and Government Sponsored Enterprises of the United States House of Representatives Committee on Financial Services*. Washington: United States House of Representatives

Scholtus, van Dijk and Fijns (2014): Scholtus M., van Dijk D., Frijns B. (2014). 'Speed, algorithmic trading, and market quality around macroeconomic news announcements'. *Journal of Banking and Finance* 38 January. pp. 89-105. doi:10.1016/j.jbankfin.2013.09.016

SEC (2010): Securities and Exchange Commission (2010). 'Proposed Rules, Concept Release on Equity Market Structure'. *Federal Register* 75(13) January

SEC (2014): Securities and Exchange Commission (2014). *Staff of the Division of Trading Markets; Equity Market Structure Literature Review – Part II: High Frequency Trading*. Washington: Securities and Exchange Commission

Serbera and Paumard (2016): Serbera JP., Paumard P. (2016). 'The fall of high-frequency trading: A survey of competition and profits'. *Research in International Business and Finance* 36 January. pp. 271–287. doi:10.1016/j.ribaf.2015.09.021

Sheskin (2004): Sheskin D. (2004). *Handbook of parametric and nonparametric statistical procedures* 3rd ed. Boca Raton: CRC Press LLC

Shleifer and Vishny (1997): Shleifer A., Vishny R. (1997). 'The Limits of Arbitrage'. *The Journal of Finance* 52(1) March. pp. 35-55

Shostak (1997): Shostak F. (1997). 'In Defense of Fundamental Analysis: A Critique of the Efficient Market Hypothesis'. *Review of Austrian Economics* 10(2). pp. 27-45

Sornette (2003): Sornette D. (2003). 'Critical market crashes'. *Physics Reports* 378. pp. 1-98

Sornette and von der Becke (2011): Sornette D., von der Becke S. (2011). 'Crashes and High Frequency Trading'. *Foresight Driver Review DR7*. UK Government Office for Science

Steiner (2010): Steiner C. (2010). *Wall Street's Speed War*. Available at: www.forbes.com. Accessed on 27/6/2014

Stiglitz (2014): Stiglitz J. (2014). *Tapping the Brakes: Are Less Active Markets Safer and Better for the Economy?* 2014 Financial Markets Conference. 15/4/2014. Atlanta. Federal Reserve Bank of

Atlanta

Stoll and Whaley (1990): Stoll H., Whaley R. (1990). 'The Dynamics of Stock Index and Stock Index Futures Returns'. *Journal of Financial and Quantitative Analysis.* pp. 441–468

Subrahmanyam (2013): Subrahmanyam A. (2013). 'Algorithmic trading, the Flash Crash, and coordinated circuit breakers'. *Borsa Istanbul Review* 13. pp. 4-9. doi.org/10.1016/j.bir.2013.10.003

Taleb (2007): Taleb N. (2007). *The Black Swan*. New York: Random House Inc.

Trivedi (2016): Trivedi K. (2016). *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. Hoboken: John Wiley & Sons Inc.

Van Horne (1995): Van Horne J. (1995). *Financial Management and Policy*. Englewood Cliffs: Prentice Hall International

van Kervel (2015): van Kervel V. (2015). 'Competition for Order Flow with Fast and Slow Traders'. *Review of Financial Studies* 28(7). pp. 2094-2127. doi:10.1093/rfs/hhv023

Vella and Ng (2016): Vella V., Ng W.L. (2016). 'Improving risk-adjusted performance in high frequency trading using interval type-2 fuzzy logic'. *Expert Systems With Applications* 55. pp. 70–86. doi:10.1016/j.eswa.2016.01.056

Vuorenmaa and Wang (2014): Vuorenmaa T., Wang L. (2014). *An Agent-Based Model of the Flash Crash of May 6, 2010, with Policy Implications*. Available at: ssrn.com/abstract=2336772. Accessed on 21/04/2014

Wah and Wellman (2013): Wah E., Wellman M. (2013). *Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model*. Proceedings article presented at EC’13. 16–20/6/2013. Philadelphia. National Science Foundation

Wahab, Lashgari and Cohn (1993): Wahab M., Lashgari M., Cohn R. (1993). 'Arbitrage Opportunities in the American Depository Receipts Market Revisited'. *Journal of International Financial Markets, Institutions & Money* 2(3-4).

Weaver (2012): Weaver D. (2012). 'Minimum obligations of market makers'. *Foresight Driver Review EIA8*. UK Government Office for Science

Wilson and Marashdeh (2007): Wilson E., Marashdeh H. (2007). 'Are Co-integrated Stock Prices Consistent with the Efficient Market Hypothesis?'. *The Economic Record* 83(S1). pp. S87-S93

Yilmaz et al. (2015): Yılmaz M., Erdem O., Eraslan V., Arık E. (2015). 'Technology upgrades in emerging equity markets: Effects on liquidity and trading activity'. *Finance Research Letters* 14. pp. 87–92. doi:10.1016/j.frl.2015.05.012

Zervoudakis et al. (2012): Zervoudakis F., Lawrence D., Gontikas G., Al Merey M. (2012). *Perspectives on High-Frequency Trading*. London: University College London

Zhang (2010): Zhang F. (2010). *High-Frequency Trading, Stock Volatility, and Price Discovery*. Available at ssrn.com/abstract=1691679. Accessed on 31/07/2014. doi:10.2139/ssrn.1691679

Zhang and Baden Powell (2011): Zhang F., Baden Powell S. (2011). 'The Impact of High-

Frequency Trading on Markets'. *CFA Institute Magazine* 22(2) March-April. pp. 10-11. doi:10.2469/cfm.v22.n2.3

Zigrand, Cliff and Hendershott (2012): Zingrand JP., Cliff D., Hendershott T. (2012). 'Financial stability and computer based trading'. *Foresight Driver Review WP2*. UK Government Office for Science