The Impact of Quote-Stuffing in High-frequency Trading

Abstract

This paper discusses high-frequency trading and how quote-stuffing could affect the function of trading systems. A financial market simulator has been developed to demonstrate how quote-stuffing can increase the gap of best bid and ask prices between financial markets. This has some very important implications as traders could profit by artificially creating latencies in trading data feeds, taking advantage of induced price differences between financial markets. Recommendations are put forward in relation to how efficient market monitoring and surveillance systems and order control mechanisms could be implemented in order to safeguard the proper behaviour of trading system.

Keywords

high-frequency trading, flash crash, quote-stuffing, financial markets, market manipulation, fraud detection, market simulator, market monitoring, market surveillance, complex event processor

*A previous version of the paper was submitted to BALAS 2015, however, although accepted, the paper was not presented at the conference. This revised version includes a new implementation in JAVA environment, updated literature review and enhanced description of the simulation parameters. This paper has not been published by any other Journal or Conference.

1 Introduction

The general goal of this research is to study different configurations of trading and new potential types of market manipulation activities in order to help create, test, and continuously enhance surveillance and market monitoring systems. Generally speaking, market monitoring and surveillance is a task for which exchanges have front line responsibility for whereas regulatory authorities may initiate an investigation at a later stage based on requests from the exchanges or their own teams of analysts. The generic description of a monitoring system is that it is a highly computerized surveillance system, which alerts specific staff of unusual trading activity based on orders and executed trades (Díaz, Theodoulidis, & Sampaio, 2011).

Against common beliefs, current market monitoring and surveillance systems performance is far less efficient that one could expect. According to previous studies, for example, self-regulating exchanges carry out surveillance of on average fourteen or fifteen types of manipulations on a single-market basis, and only on two or three cross-market type of manipulations (Cumming & Johan, 2008). Evidence indicate that exchanges rank their surveillance efforts very poorly, with an average score of 2.15 out 5 for effectiveness. In fact, the proportion of actual manipulations detected in relation to the number of trades (excluding false positives) is in the order of 1.29% for self-regulating exchanges and 0.61% for regulators (2005 figures). This compares closely with the figures reported for other markets, who estimates that approximately 1.1% of closing prices are manipulated (Putnins, 2009). Furthermore, for every prosecuted closing price manipulation, there are approximately three hundred instances of manipulation that remain undetected or are not prosecuted (2009 figures). In this scenario this research posits a review of existing surveillance and monitoring techniques and systems, in order to propose a revised set of methodologies and technologies that take into account the settings in which trading takes place in today markets.

In particular, one area that is a matter of controversy is the role of Algorithmic Trading and the new High Frequency Trading (HFT) platforms, and whether it is possible to manipulate the markets using these technologies in such a way that current Financial Markets Monitoring and Surveillance Systems, FMMSS here on, will not be able to deal with them. The specific goal of this work is to study the potential consequences of ill-intended HFT and HFT-like practices, with an emphasis on how the interplay between low- and high-frequency trading could affect markets and FMMSS performance.

Given the fact that is very difficult to find complete data on confirmed quote-stuffing cases, a market simulator is introduced. The simulator was built and used to model trading systems´ behaviour when confronted with HFT quote-stuffing practices. In particular, the simulator was used to analyse the process of sending, receiving and processing orders originating from HFT and HFT-like agents, studying the effects of quote-stuffing orders on the display book, including best bid and ask prices, as well as trade prices and volumes. Of especial interest was also the monitoring of the trading system performance with the intention of studying market monitoring and surveillance systems reliability and resilience.

These initial set of simulations presents preliminary evidence on how quote-stuffing could increase the gap of prices between markets, contrary to the evidence that HFT has helped align prices across markets (Hendershott et al. 2011) and that HFT may have economically negligible effects for retail investors (Ende, Uhle, & Weber, 2011). Using these findings, a set of recommendations are discussed to exemplify how monitoring systems and efficient circuit breaker mechanisms and order control mechanisms could be implemented by an exchange in order to safeguard appropriate behaviour in its trading system. Finally, suggestions are put forward on how a market monitoring and surveillance engine could have been used to detect potentially manipulative HFT behaviour and other misconducts.

2 Background and Literature Review

2.1 Algorithmic Trading and HFT

The latest technological development used in securities' trading is called High-frequency Trading (HFT), which refers to the use of computer algorithms to generate and submit orders to electronic markets at high speed i.e., thousands of orders within milliseconds (Avellaneda & Stoikov, 2008). These algorithms take into consideration various factors to determine how, where and when to trade. These include the price of a security, the size or liquidity available, various timing considerations (e.g. how quickly can an order be executed or when exactly orders should be placed to ensure the biggest chance of execution), how likely an order is to be filled (the "fill ratio") and the overall monetary costs of each transaction ("High-frequency Trading & Algorithmic Trading," 2011).

An important component of an HFT algorithm is the trading strategy that is followed. Generally speaking, those strategies monitor the markets and their products for imbalances in prices in order to decide when to raise or cancel an order. When alerts are raised, a massive volume of orders is generated targeting at taking advantage of those imbalances (Chlistalla, 2011). Currently, HFT signifies as much as 50% of the volume traded in the USA (U.S. Securities and Exchange Commission, 2014), but as much as 90% of those orders are cancelled within milliseconds (Nanex, 2011). Most HFT firms are proprietary firms, and consequently, risk their capital to provide liquidity. Hence, HFT firms typically avoid holding positions in a particular security for long periods. In practice, HFT firms have become unofficial market markers with none of the traditional market-maker obligations (Brogaard, 2010). In simple terms, even though HFT firms are willing to offer a buy or sell side at a given price, the amount of shares offered at this price will be typically small and available only for a few milliseconds.

A recent summary of the economic literature relating to HFT by SEC (U.S. Securities and Exchange Commission, 2014) highlighted the challenge facing any researcher which relates to obtaining useful data that can identify HFT activity. Publicly available data on orders and trades does not reveal the identity of buyers and sellers. As a result, it is impossible to identify orders and trades as originating from an HFT account when relying solely on publicly available information. In the SEC literature review, 31 papers that used non-public information were reviewed as these could identify, to a greater or lesser extent, the complete activity arising from HFT accounts. The papers used four different types of datasets: data for equity trading on NASDAQ, data on trading in the E-Mini, data that was used by CFTC and SEC staff to prepare their report on the Flash Crash and finally, a variety of datasets made available to researchers by exchanges and regulators internationally. The SEC literature review is the most comprehensive currently and it concludes that the current literature does not reveal a great deal about the extent or effect of the HFT arbitrage strategies and structural strategies because the HFT datasets generally have been limited to particular markets or products and thus, they provide little opportunity to assess HFT strategies that simultaneously seek to capture price differentials across different products and markets.

2.2 Market Manipulation and HFT

The interplay between low- and high-frequency trading in relation to their effects on asset pricing dynamics has been also studied and it has been argued that the presence of high-frequency trading increases market volatility and plays a fundamental role in the generation of flash crashes (Leal, Napoletano, Roventini, & Fagiolo, 2014). In contrast, it has been argued that reducing the latency can have positive effects on liquidity and price discovery. On April 2007, Deutsche Boerse made an important upgrade to their trading system and latency was reduced from 50ms to 10ms. As a result of this, both quoted and effective spreads decreased and the contribution of quotes to price discovery doubled to 90% post upgrade, indicating that prices are more efficient (Riordan & Storkenmaier, 2012).

Trying to understand this contrasting evidence, the Concept Release (U.S. Securities and Exchange Commission, 2010) asked market participants whether HFT directional strategies – order anticipation and momentum ignition – "may pose particular problems for long-term investors" and "may present serious problems in today's market structure". As such, momentum ignition strategies can be seen as a particular type of manipulation episode as it is defined as initiating a series of orders and trades in an attempt to ignite rapid price move up or down (U.S. Securities and Exchange Commission, 2014). These momentum ignition strategies have also been termed "quote-stuffing" as they comprise episodes in which thousands of trading messages or "quotes" are send to the markets with no other intention than saturate market systems or confuse other traders or trading systems. The Concept Release presented also the question whether additional regulatory tools were needed to address ill-intended practices associated with momentum ignition strategies.

The complex settings of financial markets in which low and high frequency trading interact in now, not only, cross-market, but also, cross-border trading is making the task of monitoring and surveillance even more challenging: Following the “Flash Crash”, after five months of analyses, a CFTC-SEC join investigation issued its final official report blaming the crash on an Mutual Fund using algorithmic trading and stating that was the result of a liquidity crisis (U.S. Commodity Futures Trading Commission & U.S. Securities and Exchange Commission, 2010). However, a sub-sequent independent investigation (Easley, López de Prado, & O’Hara, 2011) posed that liquidity crisis are structural part of our new high frequency world, and that order imbalances, created by human-traders that generated one-side orders in high volume with the intention of scaring liquidity providers of the opposite sides, were the ones to blame. Nonetheless, those researchers did not have access to the data that identifies the person that generated each order, and consequently could not identify specific individuals.

Five years after the first official ruling, on April 2015, and using full access to data with person identifiers, the US. Department of Justice announced that it pressed criminal charges against a Mr. Sarao a human trader, stating that “his conduct was at least significantly responsible for the order imbalance that in turn was one of the conditions that led to the flash crash” (Miedema & Lynch, 2015). Sarao allegedly used an automated program to generate large sell orders that pushed down prices. He then cancelled those trades and bought the contracts at the lower prices to benefit when the market recovered. This back-pedalling on the ruling of the regulators, which has been mentioned in the specialized press as a “Regulatory Fiasco” (Bailey & Borwein, 2015), is nothing more than an example of how much we do not understand about the new market structures, and whether the current FMMSS and regulations are up to the level they are needed to safeguard the proper behavior of the financial markets.

2.3 Order Protection Rule and HFT manipulation

In USA, the Securities and Exchange Commission has adopted a number of rules under Regulation NMS for disseminating market information (U.S. Securities and Exchange Commission, 2005). More specifically, the Regulation NMS includes rules that are designed to strengthen the regulatory structure of the U.S. equity markets including the Order Protection Rule that requires exchanges to establish, maintain, and enforce written policies and procedures reasonably designed to prevent the execution of trades at prices inferior to protected quotations displayed by other exchanges. To be protected, a quotation must be immediately and automatically accessible for all NMS stocks, which include those on the major stock exchanges as well as many over-the-counter (OTC) stocks.

In order to fulfil the requirements of the Order Protection rule, market participants rely heavily on the National Best Bid and Offer prices tape. This is a consolidated database that collects information about the best prices available in the markets at any given moment. If delays are artificially introduced to market systems, then the accomplishment of the rule by exchanges and broker-dealers becomes almost impossible because participants may think that they are trading with the latest and most up-to-date information when in reality has been delayed or contaminated by quote prices that are just noise.

One possibility of introducing delays or noise is to employ HFT, submitting and cancelling orders in a way that the trading systems become overloaded and cannot process transactions efficiently. Furthermore, market integration allow big investors to stop trading in Over-the-Counter venues at any moment and return to trading in traditional markets if they consider it more appropriate. This means that traditional markets have to be ready not only for 'lit venues' trading, but also have to be prepared for large trading volumes originating from 'dark pools' without prior warning.

3 Simulator Design and Methodology

Given the fact that is very difficult to find complete data on confirmed HFT and HFT-like manipulation cases, this research proposes the construction of a market simulator. The simulator will be built and used to model trading systems´ behaviour when confronted with HFT and HFT-like manipulation practices. In particular, the simulator will be used to analyse the process of sending, receiving and processing orders originating from low- and high-frequency trading agents, studying the effects of ill-intended orders on the display book, including best bid and ask prices, as well as trade prices, liquidity and volumes. As mentioned, of especial interest is also the monitoring of the trading system performance variables such as fulfilment ratios, speed of trading, and delays in processing with the intention of studying market monitoring and surveillance systems reliability and resilience.

3.1 Overall Design

The market simulator assumes the existence of "m" interconnected exchanges, each one receiving and processing trading messages for one security originating from traders and other exchanges. Correspondingly, each exchange process and receives trading messages from "z" traders each one using a specific trading strategy to generate orders and decide when to submit them for processing. Also, every market has its own Order Book, which in turn are synced with a central or *"National Order Book"* that collects and consolidate information about the best prices available in the markets at any given moment forming a "*National Best Bid and Offer prices tape"*. Each market can send feeds of trading announcements and order books to *"v"* Data Vendors, which can feed this information back to traders and other exchanges. The simulator assumes that each process, communication channel and link between agents in the simulated scenarios have different delay times. Figure 1 presents an instantiations of the generic design considering: 4 interconnected exchanges, 3 traders trading on each exchange, and 1 Data Vendor. As the latter receives information of trading announcements and the latest updates on individual order books, “Data Vendor” can also be considered a special kind of *“National Best Bid and Offer Prices”* tape or database.

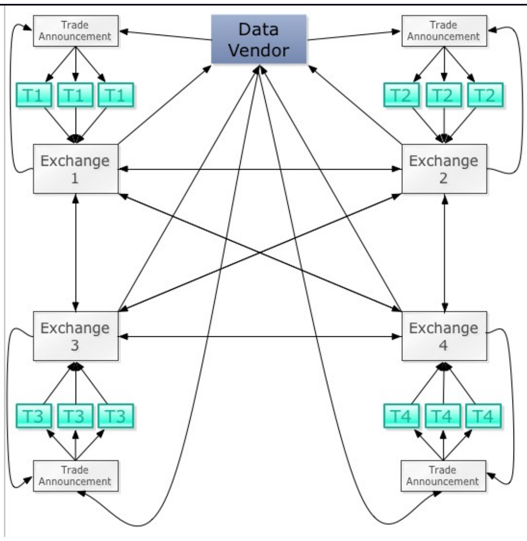

Figure 1: Simulator Design. Generic Instantiation.

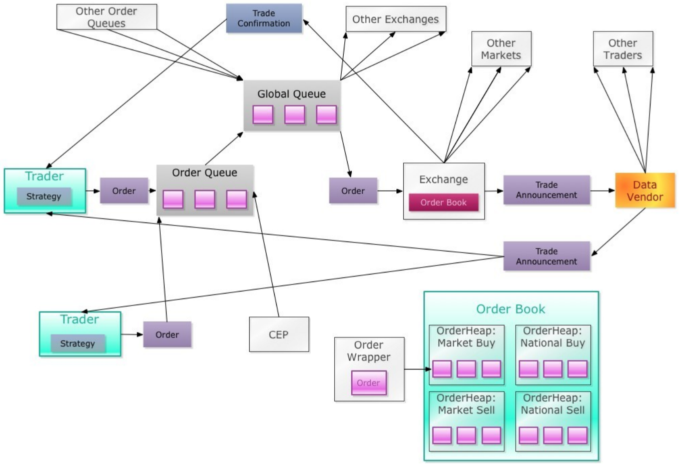

Figure 2: Trader-Exchange Simulation Design.

One major challenge was to simulate program´s execution in a parallel environment without using processing threads or other parallel processing techniques. The parallel processing environment in necessary since in real conditions each trader is free to act simultaneously with another trader. To deal with this specific challenge a *Global Queue* collects the orders from every order queue inside exchanges and prioritize them according to their arrival time. Figure 2 shows a graphic representation of the Trader-Exchange Simulation Design.

Inside each Exchange the simulator is able to execute more than one hundred different trading rules and four main types of events: *i)* messages quotes to buy or sell; *ii)* cancellation of orders; *iii)* corrections of orders; and *iv) new best* messages sent by other exchanges in order to disseminate changes in the order books. As mentioned, each trader uses a specific strategy to generate orders and submits them to the *Order Queue* of the Exchange. Exchange retrieves trading messages from the Order Queue and performs filtering and monitoring tasks using *Complex Event Processors (CEP).* If messages are considered to be in compliance with market rules and regulations then they are routed to the *Order Book*. The Order Book receives and stores valid trading messages and ranks them according to quoted prices and quantities. As mentioned, Exchanges are linked with a National Best Bid and Offer Prices databases and use this information to rank trading orders according to its internal priority but also in relation to other national markets. If the coming order cannot be executed in the specific market, then it checks in the national order book where the best available price for every share is stored. If the order can be executed in another market then the first market sends the order to the market that holds the best available price. When the order is executed, that market sends *Trade Announcements* to its Traders and to the *Data Vendor*. Then, the Exchange that made the trade sends a trade announcement to every market in order to update their national order book. Furthermore, traders accept and buffer trade confirmations from Exchange and trade announcements from Data Vendor, keeping their own orders and trade records history.

In Figure 3, Panel A it is possible to appreciate some of the parameters available at the simulator. In particular, it is possible to see two generic Exchanges: M1 and M2. Each Exchange takes an amount of β_m milliseconds to process trading messages originating from two Traders T1 and T2; $\alpha_{z,m}$ is the time that takes a message sent by Trader *z* to arrive at Market *m;* $\varepsilon_{m,z}$ is the time that takes trading confirmation and other information to travel in the opposite direction. *Time Step* or TS_z represent the pace at which Traders submit messages to each Exchange. This time is also equivalent to the difference in time between two messages or Orders (O) sent by the same trader: $TS_z = (O_t^M - O_{t-1}^M)$. Each Exchange submits trading information about prices, executions, cancellations, and other events to a *National Best Bid and Offer prices tape* or National Book (NB) following a certain periodicity, or *Time Between Updates*. Similarly, the National Book disseminate information back to Exchanges following the same logic. The time between updates is a function of TS_z and two constants i_m and j_m. The first constant represents a delay in Exchange *m* to submit information to the NB. The second constant represents a delay from NB to submit information to Exchange *m*. The total travelling time of information coming to and from each Exchange towards the NB are calculated as P_m and R_m which incorporate a component for random delays. Particularly, $P_m = i_m \cdot TS_z + \gamma$ and $R_m = j_m \cdot TS_z + \delta$, where γ and δ represent random shocks of delay. Moreover, α, β, γ, δ and ε are defined as random variables that follow a given probability distribution. The current implementation supports Normal, Fisher and Uniform distributions forms. In Figure 3, Panel B it is possible to appreciate how the different parameters define the moment at which different events take place in the time line. For instance, given TS, Trader 1 sends and order to Exchange 1 at moment t_0, given $\alpha_{1,1}$ the message arrives at Exchange 1 at moment t_1 and so on. Finally, once the message has been processed and received at the NB, the confirmation message and updated market information reach back at Trader 1 at moment t_5 as a function of the realization of the set of parameters and random variables *i, j, α, β, γ, δ* and *ε*. More details in relation to order and trading messages resolutions is presented in the next section.

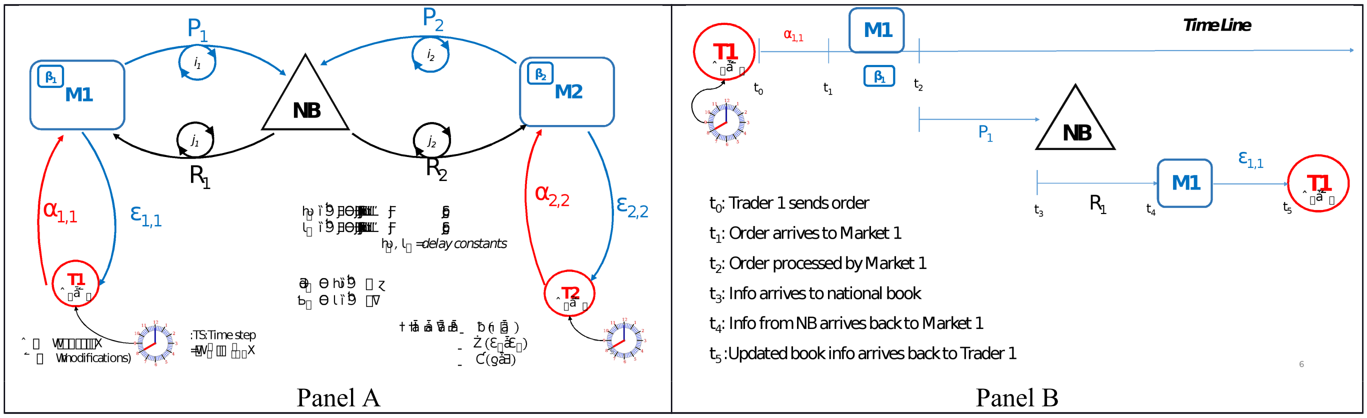
P1
P2
M1
M2
NB
β1
β2
i1
i2
j1
j2
R1
R2
α1,1
ε1,1
α2,2
ε2,2
T1
T2
delay constants
:TS:Time step
modifications)
Panel A
Time Line
α1,1
t0
t1
t2
t3
t4
t5
M1
β1
P1
NB
R1
ε1,1
t0: Trader 1 sends order
t1: Order arrives to Market 1
t2: Order processed by Market 1
t3: Info arrives to national book
t4: Info from NB arrives back to Market 1
t5 :Updated book info arrives back to Trader 1
Panel B

Figure 3: Parameters available in the Simulator

3.2 Simulation Process and Timestamp Considerations

In each simulation cycle each trader is allowed to place an order or trading message in the order queue. These orders should be executed according to their effective resolution time. Suppose that Trader A, which is assigned to the Exchange A, sends an order with a timestamp A > timestamp B and the Trader C, which is assigned to the Exchange C sends an order with a timestamp B < timestamp A. Although the Trader A's order entered first in the simulator and the Trader's C second, the order that should be executed first is the order from the Trader C. In the while, if an order or a message arrive from another Exchange or Trader with a timestamp X where A>X>B then the execution order of these three orders / messages will be 1. B -> 2. X -> 3. A

This issue arises because orders resolution from other exchanges might directly affect the resolution in other integrated exchanges, for instance, when modifying the Best Bid and Ask Prices available at the time of the effective resolution of the given order. To ensure effective resolution of trading messages the running time of the simulator is independent of the time measurements that are taken on the execution of orders. The running time of the simulator is continuous and it is defined before the start of the execution. The execution time of the orders is discrete and has as a starting point the timestamp that it is assigned to the order by the trader at the beginning of each simulation cycle.

Each order is time-stamped five times: The trader assigns the first time-stamp to the order. The second time-stamp is referring to the time that the order is received by the exchange. The third time-stamp is referring to the delay of the CEP process. The fourth timestamp is based on the processing time that the exchange needs in order to execute the order and the fifth time-stamp is the time of the arrival of the order back to the trader. This design responds to the research needs arising from the integration of different trading platforms in order to identify possible unfair advantages for a specific group of investors. Regarding the identification of each order in the simulator a unique ID is assigned. This ID may be random in case of simulated data or it can be the ID that is provided from real historical data. In each order's ID, the researcher adds different string patterns depending on the route that the order will follow. This design helps the user to follow the route of the order inside the system easily.

4 Case Study

4.1 Data Description

The case study uses real trading data for the ProShares Ultra Silver Exchange Trade Fund (ETF), symbol: AGQ obtained from NASDAQ on-demand data services (NASDAQ OMX Group, 2011). The dataset includes all quotes data received by the exchanges listed in Table 1 for a two-minute period. Between 14:36:00 and 14:46:00 hours on May 6, 2010, thousands of strange quotes for the ProShares Ultra Silver ETF were sent to NASDAQ with no specific reason or new information that could explain them. More specifically, in the two-minute period starting at 14:41:000 and ending at 14:42:59.999, a total of 1,766 quotes were sent to the market with an expiration life of 0 milliseconds. In the majority of these quotes, prices were outside the best bid and ask offers, i.e. effectively these were orders that had little or no chance of being executed at available prices. Moreover, at 14:41:18.000, the International Securities Exchange received an ask quote with price and quantity equal to 0. In this period, every quote had a valid quote condition ‘R’, representing a *regular* quote.

Market Centre Symbol	Market Centre Code	Market Centre Name	Quotes
ISEG	I	International Securities Exchange	764
ARCX	P	NYSE Arca	1848
NSDQT	T	The NASDAQ Stock Market LLC	3717
BATS	Z	BATS Exchange Inc.	2091
		Total	**8420**

Table 1: Exchanges for AGQ.

In total, the dataset includes 8,420 quotes, which were used as input to the market simulator for analysis. As shown in Table 2, the average life of a quote was less than 1 millisecond (0.57 ms), and some ask prices and ask quantities were as low as 0.

	Count	Mean	Standard Deviation	Maximum	Median	Minimum	Total N
Quotes_BidPrice	8420	56.97	.24	57.23	57.08	56.41	8420
Quotes_BidQuantity	8420	784.31	636.83	5200.00	800.00	100.00	8420
Quotes_AskQuantity	8420	813.91	668.43	3200.00	800.00	.00	8420
Quotes_AskPrice	8420	57.07	.67	57.65	57.18	.00	8420
Quote_Life (milliseconds)	8420	.057	.291	7.553	.001	.000	8420

Table 2: Descriptive statistics AGQ dataset.

4.2 Simulation Set-Up

The simulation considers the existence of four interconnected markets that receive and process messages for one security originating from investors and other exchanges. As the objective of the analysis is to study the effects of quote-stuffing, the main parameters in the market simulator software are the time it takes exchanges to process each of the messages and the number and type of messages that exchanges receive in a given time period. In particular, the processing time consists of a base processing time plus an additional time for events that generate *new best* messages.

One important aspect of the investigation is to analyse the effects of message cancellations on the overall performance and behaviour of the system. It was assumed that for some of the original quotes that did not meet certain criteria, it would be necessary and appropriate to cancel them. More specifically, the logic for generating cancelling messages is as follows: if the life of a quote is less than the total time of processing in a given exchange, then this quote is considered by the CEP *dead on arrival* and it is cancelled. Accordingly, a new 'cancel' message was added to the dataset for each original quote in the dataset whose quote life is less than the processing time. In particular, the time of arrival for cancelled messages was set the same as the time of arrival of the quote it was cancelling, but the sequence of arrival was set as immediately after the original quote. This meant that in the simulation software, although two messages had the same arrival time, messages were always processed one after the other following the sequence of arrival. Figure 4 presents a schematic view of the processing times for one market. In this figure, the top timeline represents the generic case in which messages are queued up while another message is being processed. The bottom timeline shows how two consecutive messages, one 'normal' quote message and its 'cancellation', arrive and are processed sequentially.

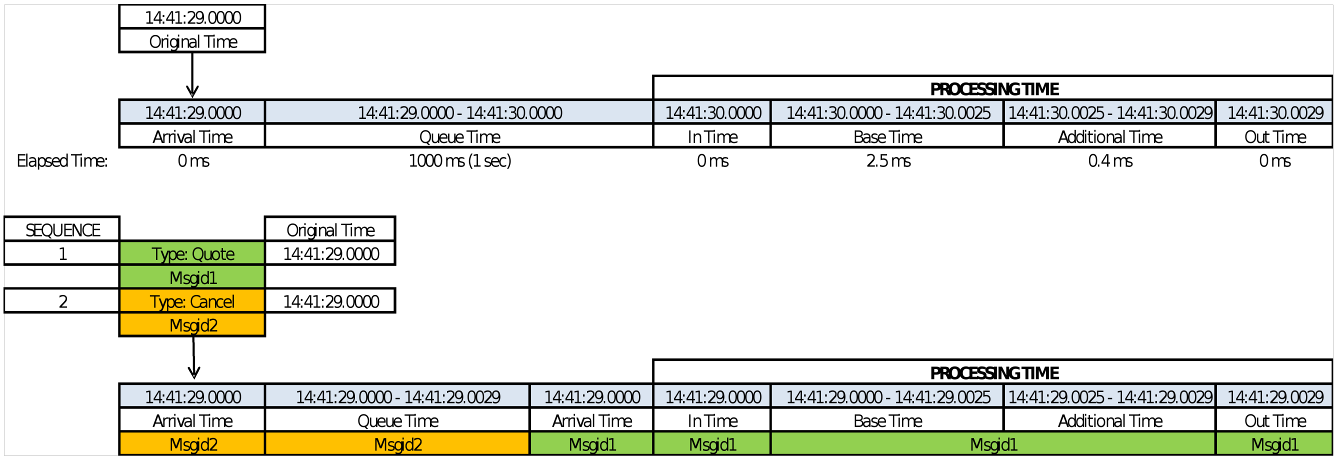

Figure 4: Simulation processing times

In total, four simulation scenarios were run, namely S1, S2, S3 and S4. For all of these, it is assumed that the order book starts with the same base prices (best bid at $57.12 and best ask at $57.13) for all exchanges; that all messages travel at 0.9 times the speed of light; that no message is corrected (in the original dataset there were no corrected messages either); and that markets are located equidistant to each other (a distance of 5 km was assumed which means that it takes 0.019ms for a message to travel from one market to another).

Table 3 summarizes the parameters for each of the simulation scenarios. In simulation S1 and S2, it is assumed that all 8,420 messages are valid, and none is cancelled. In simulation S1, the speed of processing for a single message is set to

0.5 ms with an added time of 0.4ms if the message generates a 'new best' event. In Simulation 2, the base speed is increased to 2.5ms with an added time of 0.4 ms for 'new best' events. In simulation S3 and S4, processing speeds are set the same as for simulation S1 and S2 respectively, but depending of the actual life of the quotes, some messages are cancelled immediately after submission. Due to the speed of processing in simulations S1 and S3, their exchanges are referred to as 'fast'. Simulation S2 and S4 exchanges are referred to as 'slow'. Finally, new best messages only consider the base time of processing, as these represent the dissemination of best prices from other markets, and as such, the receiving exchanges only acknowledge the arrival of the new information.

Processing Time (ms)	**S1: Fast**	**S2: Slow**	**S3: Fast**	**S4: Slow**
Quote	0.5+0.4 = 0.9	2.5+0.4 = 2.9	0.5+0.4 = 0.9	2.5+0.4 = 2.9
Cancel	0.5+0.4 = 0.9	2.5+0.4 = 2.9	0.5+0.4 = 0.9	2.5+0.4 = 2.9
New Best	0.5	2.5	0.5	2.5
Number of messages to be processed				
Quote	8420	8420	8420	8420
Cancel	0	0	3103	5825

Table 3: Simulation parameters.

In order to study the effects of the quote-stuffing in the different simulation scenarios, the best bid and ask prices are calculated as well as the gap or difference in prices that is produced at any moment in time between exchanges. To calculate the total gap in prices at timepoint *t,* i.e. $TotalGap_t,$ the squared price differences are added up for each pair of exchanges, and then the square root of the totals is calculated, as shown in Equation 1. The sum of the squared differences is used to avoid the problem of compensating differences of opposite signs.

Equation 1: Total Difference of Prices or Gap at Moment t

$$\underset{TotalGap_t=\sqrt{\text{¿}}}{\overset{BBid}{\sum_{i=1}^{4}\sum_{j=1}^{4}(\text{¿¿}i^t-BBid_j^t)^2}} + \underset{\sqrt{\text{¿}}}{\overset{BAsk}{\sum_{i=1}^{4}\sum_{j=1}^{4}(\text{¿¿}i^t-BAsk_j^t)^2}}$$

for each j>I, where $BBid_i^t$ and $BAsk_i^t$ are the best bid and ask prices at timepoint *t* for market *i.*

A cumulative and moving variable, $GapLife_{t,}$ is also calculated: this represents the duration of a gap from the timepoint it starts, i.e. the last timepoint where $TotalGap_t$ =0, until the current timepoint *t*.

Five other moving variables are defined that represent the ratio of a number of different events that generate a change in the 1st position of the order book over the last fifty (50) quotes received by an exchange. These variables measure the number of cancelling quotes, i.e. *Ratio_Cancels_Quotes*; the number of new best messages received, i.e. *Ratio_Nbest_Quotes*; the number of quotes that generated a change, i.e. *Ratio_QwCh_Quotes*; the number of re-routed messages received, i.e. *Ratio_RR_Quotes*; and the number of trades that were executed, i.e. *Ratio_Trades_Quotes*. In addition, a sixth variable is defined, *Ratio_BookChg_Quotes,* that represents the ratio of any generic event that induces a change in the 1st position of the order book over the last 50 quotes received by the exchange.

It is necessary to highlight that these variables deal only with the events that produce a change in the best position in the order book; the majority of messages did not produce this type of change. To further clarify this point, the order book keeps a record of the best bid and ask prices in descending order of priority, and as such, when the majority of orders submitted in a quote-stuffing episode are not intended to be executed, i.e. are very far away from the best bid and ask prices, they will not induce any change in the best places of the order book per se. Nonetheless, these may create delays in the refreshing processes of the order book as the exchanges must process and analyse a quote to determine that it will not produce a change, thereby increasing the length of time that other messages have to spend in the waiting queue and thus slowing down the pace of updates.

In addition, we are interested in measuring the effects of the quote-stuffing orders not only in terms of the price differences, but also in measuring the potential effects in the duration of the gap itself. Consequently, there were two working hypotheses as it was expected that quote-stuffing had a double effect on the gaps: i) they will increase the gap in terms of price differences; and ii) they will also increase its duration. Both of these effects would result from the delay introduced into the systems by the quote-stuffing orders. The different parameters used in the simulation scenarios allowed to isolate and measure these gaps under different controlled conditions, namely, the speed of processing of the exchanges, and the number of invalid or cancelled quotes that exchanges had to process during the quote-stuffing episodes. In order to measure these effects, several descriptive statistics and Pearson's correlations were calculated for each of the six gap ratios.

It is worth noticing that the input variables (ratios) are highly correlated to each other — in the best case — and that there is perfect co-linearity between some of them — in the worst case. At any single moment in time a message can only trigger a limited set of events, and that the majority of events are mutually exclusive. For instance, if at timepoint *t+m* the exchange receives a quote message, this message could actually be an original quote message or a re-routed quote message.

Besides the correlation analysis, several other descriptive statistics are calculated, including the mean, standard deviation, minimum, and maximum values of the gaps both in terms of money and time, as well as the total count of the different events. In addition, a test was run to see whether the mean gap of each simulation was statistically different to the others, in order to determine which set of parameter values created the worse (or best) conditions for trading in terms of gap duration and price differences.

5 Analysis and Results

The first step of the analysis is to examine the overall effects of the quote-stuffing in the gap and gap life. Table 4 shows the summary statistics of the gap and gap life (in seconds) for each of the simulation scenarios. It is possible to observe that both gap and gap life means are higher when the processing time at exchanges is higher (simulations with slow exchanges have greater gaps than simulations with fast exchanges), and that simulations with cancelling or dead on arrival quotes have both higher mean and higher standard deviation in terms of gap and gap life. In absolute terms, simulation S4 had the highest mean and standard deviation of gap and gap life. In contrast, simulation S1 had the smallest indicators. The differences between the means of the simulations are statistically different at the 1% level, and thus it can be concluded that these differences are not the result of chance. Figure 5 and Figure 6 show the behaviour of the gap and gap life over time in the different simulations. The time scale is set to milliseconds, starting at second 20 (20,000ms) and ending at second 120 (120,000ms).

	S1: Fast	**S2: Slow**	**S3: Fast**	**S4: Slow**
	All alive	**All alive**	**Dead quotes**	**Dead quotes**
Total Gap Mean	0.009	0.012	0.013	0.021
Total Gap St Dev	0.024	0.024	0.029	0.037
Total Gap Min	0	0	0	0
Total Gap Max	0.191	0.191	0.2	0.22
Gap Life Mean	5.609	2.655	4.322	9.036
Gap Life St Dev	7.353	3.126	5.841	10.295
Gap Life Min	0.000	0.000	0.000	0.000
Gap Life Max	26.540	13.891	22.515	33.924

Table 4: Gap and GapLife summary statistics.

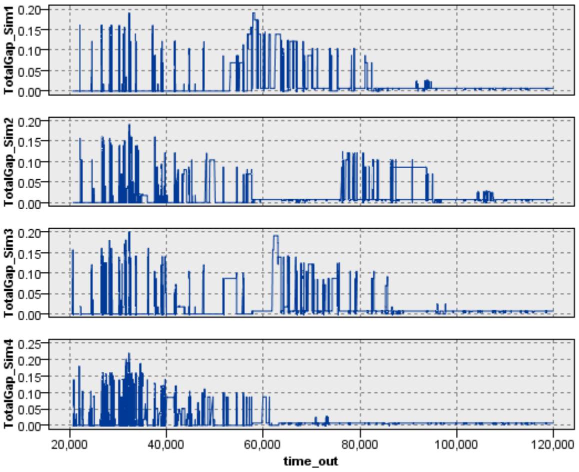

Figure 5: Total Gap

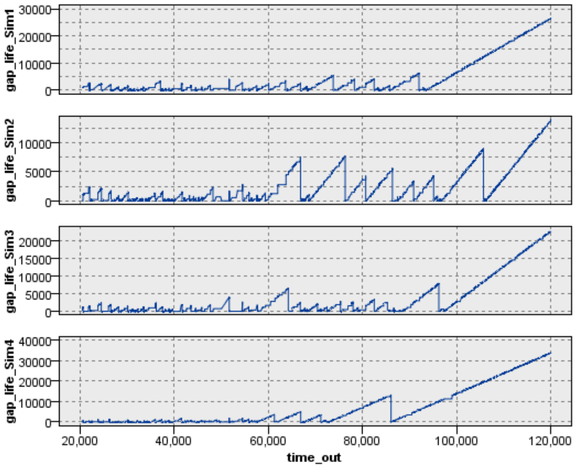

Figure 6:Total Gap Life

Table 5 shows the total number of events that were generated and processed in each simulation scenario. The count of events is broken down in terms of the total number of events and the number of those events which induced a change in the

1st position in the order book. The proportion of the latter is also given in order to facilitate comparisons. The results show how significant the effects of quote-stuffing are in putting pressure on trading systems. In simulation S1, for instance, where it is assumed that all quotes should be treated as valid, and that the processing speed is as fast as 0.9ms per message, a large number of spurious trades occur (a total of 4,862 trades).

	S1: Fast		S2: Slow		S3: Fast		S4: Slow	
	All alive		All alive		Dead quotes		Dead quotes	
	Count	% of prev. row	Count	% of prev. row	Count	% of prev. row	Count	% of prev. row
Total No of Cancels	n/a		n/a		3091		5827	
Total No of Cancels w/Chg	n/a		n/a		193	6.24%	600	10.30%
Total No of Nbest	15054		11465		17020		16388	
Total No of Nbest w/Chg	595	3.95%	1116	9.73%	1065	6.26%	1876	11.45%
Total No of Quotes	8420		8420		8420		8420	
Total No of Quotes w/Chg	299	3.55%	621	7.38%	528	6.27%	872	10.36%
Total No of RR	3117		3070		3330		3379	
Total No of RR w/Chg	297	9.53%	448	14.59%	410	12.31%	669	19.80%
Total No of Trades	4862		4732		5142		5301	
Total No of Trades w/Chg	328	6.75%	537	11.35%	460	8.95%	732	13.81%

Table 5: Gap and GapLife count of events.

Table 6 shows the results of the correlation analysis. In Table 6 Panel A, it is possible to appreciate that in general, all ratios correlate positively with the Total Gap variable, except for the new best messages ratio in simulation S1. In terms of ratios, those that contributed the most to the gaps are the Quotes with change ratio, the Cancelled quotes ratio and the New best ratio. Nonetheless, all ratios are found to be statistically significantly correlated with the Total Gap at the 1% level of significance. Table 6 Panel B shows the results of the correlation analysis this time with respect to the Gap Life. It shows that all ratios are inversely correlated with the Gap Life, which can be interpreted as demonstrating that events which did not produce a change in the best place of the order book actually contributed to extending the duration of the gap. As shown previously in Table 5, only a small proportion of messages (ranging from 3.55% to 10.36% depending on the simulation scenario) actually induced a change in the order book. Consequently, the majority of messages left the gap untouched and merely increased its duration. In terms of magnitude, it is possible to appreciate that the gap life is more closely correlated with the ratios when the exchanges are slow and when they are confronted with more cancelling messages. However, the correlation differences between simulations S1, S2 and S3 are not as significant as for the Total Gap.

Panel A	S1: Fast	S2: Slow	S3: Fast	S4: Slow
Correlations w/Total Gap*	All alive	All alive	Dead quotes	Dead quotes
Ratio_BookChg_Quotes	0.049	0.138	0.284	0.510
Ratio_Cancel_Quotes	n/a	n/a	0.301	0.458
Ratio_Nbest_Quotes	-0.017	0.094	0.278	0.474
Ratio_QwCh_Quotes	0.121	0.242	0.328	0.537

Ratio_RR_Quotes	0.077	0.032	0.071	0.265
Ratio_Trades_Quotes	0.071	0.048	0.067	0.272

Panel B
Correlations w/GapLife*

Ratio_BookChg_Quotes	-0.217	-0.227	-0.256	-0.402
Ratio_Cancel_Quotes	n/a	n/a	-0.198	-0.353
Ratio_Nbest_Quotes	-0.243	-0.205	-0.295	-0.403
Ratio_QwCh_Quotes	-0.235	-0.311	-0.260	-0.426
Ratio_RR_Quotes	-0.075	-0.034	-0.058	-0.163
Ratio_Trades_Quotes	-0.088	-0.051	-0.045	-0.163

*All correlations are statistically significant at the 1% level

Table 6: Correlation of Gaps vs Ratios.

5.1 Impact of Artificially Introduced Latency Delays

In Table 4, it is worth noticing that in the worst case, the average gap in price can be as much as $0.021 and can last for 9.036 seconds. Given that this simulation covers only two minutes of data, it is easy to see how important this can be if one considers the potential for arbitrage opportunities that could use this difference in prices to the detriment of other investors. Ten seconds is enough for ill-intended HFT traders to place and execute thousands of orders.

The results shown in Table 5 can be interpreted as illustrating that even in very fast exchanges, strange quotes can induce systems to execute a high number of spurious trades (as many as 4,862 in the two-minute period). These trades should not have occurred if a proper CEP monitoring system had had the chance to reject invalid quotes before they entered the trading systems. Considering that a quote triggers a chain of events that is almost impossible to foresee is key to understanding the 'butterfly' effect that is occurring in these simulation scenarios. If one invalid quote is treated as valid, then the way in which subsequent events and messages are resolved can be completely different to what would have happened if that invalid quote had been rejected in the first place. For instance, if an invalid quote changes the best bid in one exchange, then that mistake will affect all trading not only in that particular exchange; the error would be disseminated to other exchanges via 'new best' messages, thereby perpetuating and enhancing the original error. This holds true even if the invalid quote is later cancelled, as it was assumed in these simulation scenarios.

Table 5 should be analysed under the consideration that these figures represent the interaction of both valid and invalid messages in a simulation environment, which is different from what actually occurred during the flash crash. However, the contribution of the simulations is very straightforward because it allows to properly appreciate which conditions are worse (or better) when markets confront quote-stuffing, as well as which types of messages and conditions have the greatest impact. If one uses the number of spurious trades as an indicator of the consequences of quote-stuffing, then it is clear that both slow and fast exchanges suffer more or less the same, as the number of spurious trades is high in all the simulation scenarios. Moreover, cancelling the quotes after they have been already processed (simulations S3 and S4), actually increases the number of spurious trades, so that the effect is worse for slow exchanges.

In Table 6 Panel A, the fact that the majority of the ratios correlate positively can be interpreted as indicating that quote-stuffing episodes contain a high number of messages with the potential to misalign the best bid and ask prices in the exchanges, and that these misalignments are also disseminated to other exchanges. In terms of the magnitude of the correlations, it can be seen that in simulation S4, the correlations are higher than for the rest of the simulations. It is also evident that simulation S3 and simulation S4 both have higher correlations than simulations S1 and S2, which indicates that not only the speed of processing is important, but also the number of quotes that are later cancelled. In summary, *ceteris paribus*, the higher the speed of processing, the smaller the gap that is produced, and the higher the number of cancelled quotes, the larger the gap that is produced. This can be interpreted as signifying that quote-stuffing in general produces a number of events that are positively correlated with the Total Gap.

The results of Table 6 Panel B can also be interpreted as an indication that regardless of the speed of processing and the number of quotes that are cancelled, quote-stuffing practices produce events that extend the life of the gaps, and thus create more arbitrage opportunities for potential manipulators. Overall, the results of the simulations present evidence in favour of the working hypothesis, and thus it is possible to assert that under the controlled conditions of this case study, quote-stuffing episodes: increase the gap in terms of price differences, and they also increase its duration.

Consequently, it is possible to deduct that under these conditions, quote-stuffing episodes deteriorate market integrity, as prices in the order books do not reflect all publicly available information every time fresh data is delayed by quote-stuffing episodes. Moreover, the case study demonstrates that under controlled conditions, a higher proportion of invalid quotes, even when these are later cancelled, directly affect both the magnitude and duration of price gaps.

6 Conclusions

The purpose of this paper is to study the consequences of quote-stuffing in financial markets and the challenges that HFT has to address in order to monitor and safeguard against potential manipulative and inappropriate behaviours. In terms of the research design, a case study approach was used. The paper presented preliminary evidence on how HFT quote-stuffing could increase the gap of best bid and ask prices between markets, contrary to the previous evidence that HFT has helped align prices across markets. This has some important implications, as it is possible for a trader to profit by artificially creating latencies in trading data feeds that would make arbitrage possible by taking advantage of the HFT induced price differences between markets.

The question is, then, what steps are necessary to safeguard the spirit of rules such as the Order Protection rule, which were created to preserve market integrity, when advances in market efficiency are in direct conflict with the former? First, it is necessary to raise awareness of the unintended consequences of increasing the speed of trading, as well as to create mechanisms to guarantee that these consequences are managed and kept under control. One way to solve the problem could be to enforce a fully centralized global market, in which all trading orders, without exception, are sent and processed by one entity, which should be in charge of distributing the information to the right markets at the appropriate pace (Theodoulidis & Diaz, 2012). This would effectively introduce artificial, but controlled, delays so that all market participants have access to the same information at the same time. This solution however is almost impossible to achieve given the global interconnectivity of markets, the financial implications and perhaps more importantly, the political and governance implications of such mechanisms.

A more plausible alternative would be to enforce a policy under which exchanges are obliged to compensate investors retroactively for any losses which result from the de-synchronization of prices. This solution is similar to the clearing role markets already play. It is expected that in the absence of manipulators, the price differences would have a random, unbiased impact on both the buy and the sell sides, and thus exchanges could effectively act as trustworthy clearing parties. In order to achieve this, it will be necessary for markets to collect and store all relevant information, for instance, an order book and its depth, making exchanges accountable for any price differences that they help create or disseminate that is not properly compensated at the end of the trading day.

From the monitoring and surveillance point of view, it is clear then, that it is crucial that delays are not introduced intentionally to the systems, and thus it is necessary that exchanges are able to filter out invalid quotes before they are processed by the trading systems. With respect to the quote life, for instance, a minimum life could be enforced, much like the time-in-force policy in which all quotes are considered valid for a minimum of, for example, 50ms (U.S. Securities and Exchange Commission, 2011). With respect to the price and quantity, *collars* or *volatility interruption bands* or *circuit breakers* (Gomber et al., 2011) could be introduced at the order level (not at the trade level as it is currently done), rejecting any order which is too far from the last valid best bid or ask offer. These are collectively issues that can be addressed as part of a “Complex Event Processing (CEP)” component as it is discussed extensively in (Theodoulidis & Diaz, 2012). This would also help to tackle any stub-quoting problems i.e., when a market maker makes an offer to buy or sell a stock at a price so far away from the prevailing market that it is not intended to be executed. In addition, cancellation ratios that affect the order to trade ratios could be monitored and rejected when they exceed a certain thresholds. However, these could also be bypassed in the case where trades are executed at certain frequencies to avoid crossing the threshold (Gomber et al., 2011). An additional complication for the definition of such mechanisms, are their temporal characteristics that define their instantaneous and cumulative features i.e., the start time and duration for their enforcement.

Further research should address different ways of studying the HFT patterns of cross-border trading in real time and linking them with off-line data (Diaz & Theodoulidis, 2012). One path that could be followed is the enhancement of the market simulator with the development of additional functionality which allows the testing and checking of more types of trading rules and halts, as well as, to perform all kinds of stress tests to HFT trading algorithms and trading systems. This enhanced simulator should be tested in new case scenarios and should include functionalities for dealing with other products and securities, again with special emphasis on cross-border and non-equity markets scenarios.

7 References

Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*. doi:10.1080/14697680701381228

Bailey, D. H., & Borwein, J. M. (2015). Lessons From the “Flash Crash” Regulatory Fiasco. *http://www.huffingtonpost.com/*. Retrieved May 25, 2015, from http://www.huffingtonpost.com/david-h-bailey/lessons-from-the-flash-cr_b_7148898.html

Brogaard, J. (2010). *High Frequency Trading and Its Impact on Market Quality*. *Working Paper Series SSRN eLibrary*. Retrieved from http://ssrn.com/paper=1641387

Chlistalla, M. (2011). High-frequency trading, better than its reputation? Deutsche Bank Research. Retrieved from http://www.dbresearch.com/PROD/DBR_INTERNET_EN-PROD/PROD0000000000269468.pdf

Cumming, D., & Johan, S. (2008). Global Market Surveillance. *American Law and Economics Review*, *10*(2), 454–506. doi:10.1093/aler/ahn009

Díaz, D., Theodoulidis, B., & Sampaio, P. (2011). Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices. *Expert Systems with Applications*, *38*(10), 12757–12771. doi:10.1016/j.eswa.2011.04.066

Easley, D., López de Prado, M. M., & O’Hara, M. (2011). The Exchange of Flow Toxicity. *The Journal of Trading*, *6*(2), 8–13. doi:10.3905/jot.2011.6.2.008

Ende, B., Uhle, T., & Weber, M. (2011). The Impact of a Millisecond: Measuring Latency Effects in Securities Trading. In *Wirtschaftinformatik Proceedings 2011*.

Gomber, P., Arndt, B., Lutat, M., & Uhle, T. (2011). *High-Frequency Trading*. *SSRN eLibrary*. SSRN. doi:10.2139/ssrn.1858626

Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does Algorithmic Trading Improve Liquidity? *The Journal of Finance*, *66*(1), 1–33. doi:10.1111/j.1540-6261.2010.01624.x

High Frequency Trading & Algorithmic Trading. (2011). *HFT Review*. Retrieved August 26, 2011, from http://www.hftreview.com/pg/blog/mike/read/5307/high-frequency-trading-algorithmic-trading

Leal, S. J., Napoletano, M., Roventini, A., & Fagiolo, G. (2014). *Rock around the Clock : An Agent-Based Model of Low- and High-Frequency Trading* (pp. 1–25). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2390682

Miedema, D., & Lynch, S. (2015). UK speed trader arrested over role in 2010 U.S. "flash crash." *UK Reuters.com*. Retrieved from http://uk.reuters.com/article/2015/04/21/uk-usa-security-fraud-idUKKBN0NC21O20150421

Nanex. (2011). May 6'th 2010 Flash Crash Analysis - Continuing Developments and Research. Retrieved from http://www.nanex.net/FlashCrash/FlashCrashAnalysis.html

NASDAQ OMX Group. (2011). NASDAQ data on demand services. Retrieved from http://www.nasdaqdod.com

Putnins, T. J. (2009). Closing price manipulation and integrity of stock exchanges. *Discipline of Finance, Faculty of Economics and Business*. Sydney: University of Sydney. doi:http://hdl.handle.net/2123/5925

Riordan, R., & Storkenmaier, A. (2012). Latency, liquidity and price discovery. *Journal of Financial Markets*, *15*(4), 416–437. doi:10.1016/j.finmar.2012.05.003

Theodoulidis, B., & Diaz, D. (2012). Financial Markets and High Frequency Trading: An Information Management Perspective. *SSRN Electronic Journal*. doi:10.2139/ssrn.2178944

U.S. Commodity Futures Trading Commission, & U.S. Securities and Exchange Commission. (2010). Findings Regarding the Market Events of May 6, 2010. Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. Retrieved from http://www.sec.gov/news/studies/2010/marketevents-report.pdf

U.S. Securities and Exchange Commission. (2005). Final Rule Regulation NMS. *Release No. 34-51808* . Retrieved from http://www.sec.gov/rules/final/34-51808.pdf

U.S. Securities and Exchange Commission. (2010). Concept Release on Equity Market Structure. Washington DC. Retrieved from http://www.sec.gov/rules/concept/2010/34-61358.pdf

U.S. Securities and Exchange Commission. (2011). Order Approving Proposed Rule Changes Relating to Expanding the Pilot Rule for Trading Pauses Due to Extraordinary Market Volatility to all NMS stocks. Retrieved December 17, 2011, from http://sec.gov/rules/sro/bats/2011/34-64735.pdf

U.S. Securities and Exchange Commission. (2014). *Equity Market Structure Literature Review Part II : High Frequency Trading* (pp. 1–37). Retrieved from http://www.sec.gov/marketstructure/research/hft_lit_review_march_2014.pdf