A Model for Analyzing the Impact of High-Frequency Trading Latency On Market Quality

[1]Ebono, F.; [2]Ogunsakin, R.; [3]Ochei, L.C
Department of Computer Science
University of Port Harcourt
Port Harcourt, Nigeria.
E-mail: [1]fubara.ebono@uniport.edu.ng; [2]rotimi.ogunsakin@gmail.com; [3]laud.ochei@gmail.com

ABSTRACT

This research evaluates the impact of High Frequency Trading (HFT) Latency on market quality, which are Liquidity, Price Discovery and Volatility. To achieve this, a Mathematical model was developed to model and explain the various impact of HFTs on market quality. The Mathematical model provides an in-depth analysis of HFT Latency and its resulting impacts on market quality, which correlate with empirical results from literature and provide a causal explanation to some observed impacts of HFT Latency. The research aims to provide the foundation for developing a simulation model for understanding the impact of HFT latency on market quality, where different market parameters can be varied, such as the behaviour of market participants, to observe the impact on market quality,

Keywords: High Frequency Trading, Liquidity model, Trading model, Financial model

CISDI Journal Reference Format
Ebono, F.; Ogunsakin, R. & Ochei, L.C. (2023~~0~~): A Model For Analysing The Impact of High-Frequency Trading Latency On Market Quality. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 11 No 2, Pp 171-192
Available online at www.isteams.net.cisdijournal

1. INTRODUCTION

The arrival of automation in the exchanges gradually brings to an end the involvement of human professional floor brokers and market makers. Over time, these professional tasks were fully replaced with high-speed computers and sophisticated algorithms (Algorithms for large order processing and automated market making) and the resulting savings from high wages and commissions paid to professional human brokers and market makers are passed onto investors and traders in the form of lowered trading fee and rebate [1].

High-Frequency Trading (HFT) results from innovation in computer speed and sophisticated algorithms [2]. Technology innovation in communication, advancement in microprocessor design, and the ability to manage complex algorithms more efficiently have contributed to advancing HFT activities in recent times. HFT employs different trading strategies, which are majorly electronic liquidity provisioning, statistical arbitrage, liquidity detection and directional strategies (Market inference and News), among others. These strategies and their underlying algorithms are gradually becoming homogenous across the HFTs and have shifted competition to latency (speed) instead of strategies. This account for the high investment by “Top HFT Firms” into reducing latency from milliseconds to micro-second [1].

High-Frequency Trading (HFT) is a primary form of Algorithmic Trading (AT) characterised by the use of sophisticated technology tools and computer algorithms to rapidly trade securities [2]. According to the United States Security and Exchange Commission (SEC), High-Frequency Traders (HFTs) are proprietary trading firms that use high-speed (very low latency) systems to monitor market data, submit large amounts of orders to the market, and utilise quantitative and algorithmic methodologies to maximize the speed of their market access and trading strategies [3]. Other characteristics attributed to HFTs are (1) The use of high-speed and sophisticated computer programs to generate, route, and execute orders; (2) The use of co-location services and individual data feeds offered by exchanges and other service providers to minimise network and other types of latencies (like processing speed); (3) Very short time-frames for establishing and liquidating positions; (4) Submission of numerous orders that are cancelled shortly after submission; and (5) Ending a typical trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions overnight) [3].

HFT accounts for a market share of over 70% in the United States equity market in 2013 [4], and 40% of the Europe equity market [5], and according to TABB Group, HFT accounts for 77% of the UK equity market [2]. With the largest share of the equity market dominated by HFTs, it is evident that HFT activities would be a predominant trading activity exhibiting the highest impact on equity market quality, which are liquidity, price discovery and volatility. Thus, a detailed analysis of the impact of HFT trading activities which include trading strategies and latency arms race on market quality is highly imperative and beneficial to the equity market and the global economy.

Research works in this domain have examined the impacts of HFT strategies on market quality using trading data [5, 7, 8, 10]. There are also research works that have examined the impact of latency on market quality with a main focus on exchange latency [9] or HFT Latency [12, 13, 14] using trading data. None of the research on the effect of latency on market quality has explored a mathematical model approach to modelling these impacts based on the outcome from the different analyses to gain insight into the fierce latency competition taking place among the HFTs and their corresponding impact on equity market quality.

Therefore, in this research work, a Mathematical model is developed based on valid assumptions from empirical and theoretical results in the research literature, to analyse and evaluate the impact of HFT latency on market quality - liquidity, price discovery and volatility. The results of the mathematical model s are juxtaposed with that of the empirical results from research literatures. The mathematical model provides a foundation for simulating HFT trading and latency activities and testing novel trading and latency activities in the liquidity market.

The remaining of chapters is structured as follows. In Section 2, a literature review of the evolution of HFT as a subset of Algorithmic Trading (AT); the different HFT strategies and their impact on equity market quality; the Effect of HFT latency variation on market quality; and the application of theoretical model (Mathematical model) in the domain of financial intelligence. Section 3 contains the Research Design and Methodology, where the theoretical and mathematical model s are discussed in detail, including the underlying assumptions, theoretical logic and approach used. Section 4 contains the implementation of the Mathematical and the discussion. Section 5 includes the conclusion and recommendation for further work.

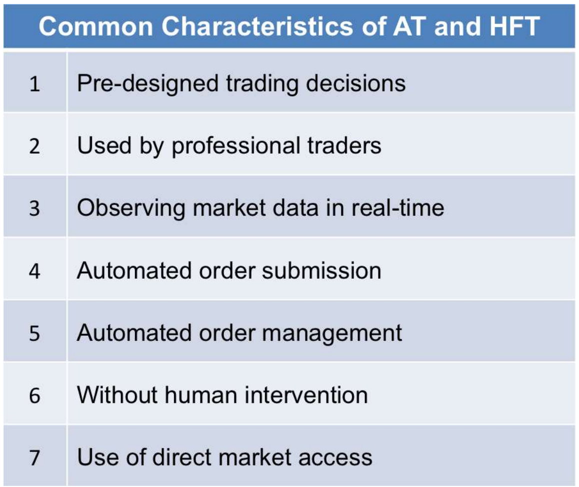
2. LITERATURE REVIEW

2.1 High Frequency Trading And Algorithmic Trading

In algorithmic trading (AT), computers are directly interfaced with trading platforms and submit orders to the exchange without immediate human intervention. These computers possess inbuilt sophisticated algorithms that make trading decisions based on the result of observed historical market data and other information at very high frequencies, often in milliseconds [18].

HFT, as a subset of Algorithmic Trading (AT), also possesses similar characteristics as AT, that is, the use of sophisticated algorithms to make trading decisions in milliseconds. However, the difference between AT and HFT is that ATs mostly have longer holding periods: minutes, days, weeks, months or even longer. Whereas HFTs hold a very short position, mostly in seconds or less and attempt to close the trading day in a neutral position[5].

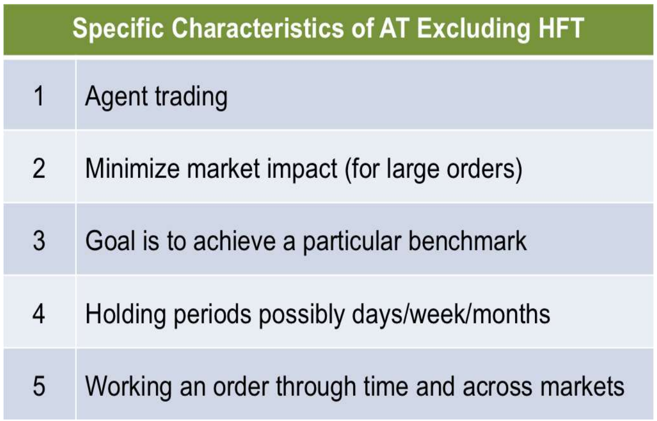
To further differentiate between ATs and HFTs, Characteristics common to AT and HFT and those specific to AT and HFT are presented (see Figure 2.1).



Common Characteristics of AT and HFT	
1	Pre-designed trading decisions
2	Used by professional traders
3	Observing market data in real-time
4	Automated order submission
5	Automated order management
6	Without human intervention
7	Use of direct market access

Figure 2.1: Common Characteristics of AT and HFT [19]

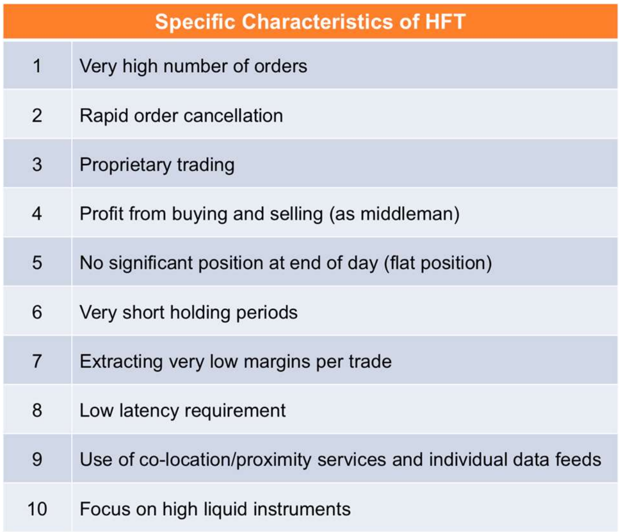
There are characteristics that are specific to AT and not commonly associated with HFT. The focus is usually on intelligently working orders through time and across markets to minimise the impact of large orders relative to a pre-defined benchmark[19]. The Table in Figure 2.2 shows those characteristics specific to AT commonly not associated with HFT.



Specific Characteristics of AT Excluding HFT	
1	Agent trading
2	Minimize market impact (for large orders)
3	Goal is to achieve a particular benchmark
4	Holding periods possibly days/week/months
5	Working an order through time and across markets

Figure 2.2: Specific Characteristics of AT Excluding HFT [19]

HFT strategies are naturally geared towards a highly liquid instrument, and as a prerequisite, HFTs rely on high-speed (low latency) access to markets, achieved by high investment in high-speed communication linked to the exchange, usage of co-location/proximity service, and dedicated/individual data feed. The Table in Figure 2.3 shows characteristics specific to HFT, which are usually not associated with AT.



Specific Characteristics of HFT	
1	Very high number of orders
2	Rapid order cancellation
3	Proprietary trading
4	Profit from buying and selling (as middleman)
5	No significant position at end of day (flat position)
6	Very short holding periods
7	Extracting very low margins per trade
8	Low latency requirement
9	Use of co-location/proximity services and individual data feeds
10	Focus on high liquid instruments

Figure 2.3: Specific Characteristics of HFT [19]

2.2 High Frequency Trading (HFT) Strategies

While the diversity and opaqueness of the HFT universe make it difficult to be able to examine all strategies, there are well-known strategies, most of which were in existence before the advent of HFT but were made more effective through the use of high-speed computing infrastructures, communication networks, and sophisticated algorithms. Figure 2.4 below shows the list of some of the popular HFT Strategies as identified in the research domain.

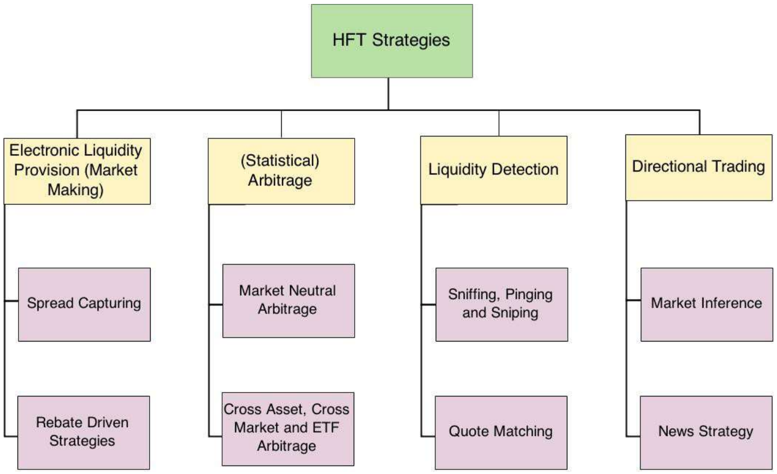


Figure 2.4: Common High Frequency Based Strategies - Adapted from [19]

Electronic Liquidity Provision (Market Making)

Market making refers to a strategy in which a market maker (in this case, an HFT Trader) simultaneously posts both buy and sell limit orders for a financial instrument on both sides of the electronic order book in order to profit from the bid-ask spread and at the same time providing liquidity to market participants [6, 19]. To encourage the provision of liquidity in the equity market, HFTs who provide liquidity are given rebates as compensation for the risk involved in liquidity provisioning, thereby ensuring all market orders are converted to trade at the market bid-ask price at any given time during the trading hours.

(Statistical) Arbitrage

Statistical Arbitrage refers to a strategy in which two closely related financial instruments whose prices should move on-a-pair with each other, for example, the S&P 500 features and SPY (the ticker symbol that tracks S&P 500). If the price of S&P 500 goes up due to the arrival of a buy order and the SPY do not go up immediately, an HFT can buy the low-price SPY and sell the high-price S&P 500, profiting from the bid-ask spread [6]. Opportunity for this type of strategy only exists in the market for a very short period (in micro or milliseconds), and thus the fastest HFTs will always win it all. Other forms of arbitrage exist and predominantly benefits from price inefficiencies either across asset or across market [19].

Liquidity Detection

This is a type of strategy in which HFTs employs sophisticated algorithms to discern patterns left by other market participants in the market and adjust their actions (to buy or to sell) accordingly. HFTs, most time, focus attention on large orders and employ various strategies to detect sliced orders or gain information about electronic order books - This is sometimes referred to as “sniffing out” other algorithms or “Ping” in order books to retrieve information [19].

Directional Trading

This is a type of strategy where HFTs placed orders based on order flow signals. This can also be in form of news, where news is automatically parsed using text analytics and trade (buy or sell financial instruments) based on knowledge inferred from the parsed news [6].

2.3 Impact of HFT Strategies On Market Quality

A study conducted on the New York Stock Exchange (NYSE) automated quote dissemination in 2003, using the change in the market structure that led to an increase in Algorithmic Trading (AT) activity to measure the causal effect of HFT on liquidity [7]. The study concluded that for large stocks, HFT narrows spreads, reduces adverse selection and reduces trade-related price discovery. In general, they concluded that HFT improves liquidity and enhances the informativeness of quotes.

The study of [11] in 2007 on the entry of HFT into the trading of Dutch Stocks found that the market makers (AT and HFT) inventory revert to the mean position numerous times within a trading day, which implies the presence of a high rate of liquidity within the market compared to the Belgian Stock Market (Used as a control since there was no AT or HFT presence). In comparison to the Belgian stock Market, the bid-ask spread was 15% narrower, the adverse selection was 25% less, and volatility was unaffected. This implies that the presence of HFT increases liquidity, reduces adverse selection and has no observable influence on volatility. The research from [11] shows that bid-ask spread and volatility are not correlated, contrary to theoretical observations. The contribution of HFT and non-HFT liquidity supply and liquidity demand to price change components was measured by [20], and found out that HFTs trade in the opposite direction to the market price movement. When prices deviate from fundamental value, HFTs push prices back to their efficient level by initiating a trade in the opposite direction. Contributing to price efficiency and increasing liquidity.

Consequently, [21] analysed trading equilibrium for a given level of HFT and discovered that when some HFT becomes very fast, it leads to increased adverse selection cost for the slower traders and generate negative externalities. This is probably due to the ability of the faster HFTs to process bid-ask quote information and adjust their trading strategies accordingly before the slower HFTs, leading to adverse selection for the slower HFTs. [22] studied the effect of HFT intensity on market liquidity, short-term volatility, and price discovery between 2001 and 2011 in 42 equity market around the world and found out that on average - HFT improves liquidity and price discovery but increases volatility. In contrast to the average effect, they discovered that increase in HFT intensity reduces liquidity and increase volatility for the small cap stocks. The reason for this variation was not explicit in the literature. This research work will attempt to provide a causal explanation to the resulting effect of HFT intensity (HFT intensity can also be attributed to low latency) on market quality.

The flash crash event that occurred on May 6, 2010, is one of the major externalities attributed to HFT activities. But Kirilenko et al. [8] empirically showed that HFTs did not cause the Flash Crash, but contributed to it by demanding immediacy ahead of other market participants. This immediacy absorption activity of HFTs results in price adjustments that are costly to the slower HFTs and the traditional market makers, thereby resulting in adverse selection. They also observed that at times of market stress and high volatility, which might be due to a large buy or sell order resulting in an order flow imbalance. HFTs exacerbate this directional move by demanding liquidity in the direction of price movement, increasing the speed at which the best bid and ask queue gets depleted, leading to a spike in trading volume and setting the stage for a flash-crash-type event.

In summary, most of the research showed that HFT Strategies contributed to increased liquidity in the equity market but, at the same time, generated adverse selection for the slower HFTs leading to negative externalities. Improvement in liquidity is attributed to the reduction in the bid-ask spread and volatility, but the observation by [11] shows that bid-ask spread and liquidity are not correlated. Other empirical results supported observation by [11], such as [22], which shows that HFT improves liquidity and price discovery but increases volatility. The effect of HFT on volatility seems to be inconsistent consistent in the research literature, but in general, most of the research agreed that HFTs increase volatility but also contributed to stabilising the market bid-ask quote under extreme price imbalance.

The contribution of HFT to price discovery is consistent in both theoretical and empirical research. HFT improves price discovery for the faster HFTs but worsens for the slower HFTs leading to adverse selection for the slower and non-HFTs. The causal explanation for this is provided in the developed Mathematical model .

2.4 High Frequency Trading (HFT) And Latency

The assumed speed limit for the trading world is said to be the speed of light which is almost impossible to achieve considering the natural limitation set by the physical component used in transmitting market data from an exchange to a trader and back to the exchange. Even in the case of co-location, where HFTs systems are connected directly to the exchange, the connection medium still poses a barrier to achieving this speed of light. The clock speed of the hardware on which trading algorithms are executed is also another barrier.

Despite these natural barriers to achieving the speed of light in the trading world, where trading decisions are made and executed in milli-seconds, such spheres of operations have no space for human traders [23]. Latency “The time it takes to access, process and respond to market information [24]” is a relative term for HFTs, since yesterday’s ultra-low latency can be today’s low latency. However, low latency can be classified as those with sub-ten (single digit) milli-seconds round trip [23]. In the exchange market, latency exists in two forms which are Exchange latency and Market players latency which are the HFTs (Agency and Proprietary), non-HFTs and Human traders. These two forms of latency exhibit a considerable impact on market quality.

Exchange And Latency

The business model of exchanges is such that its functionality depends on its ability to receive, aggregate, manage, match orders and generate resulting trade information. Contrary to the old traditional exchange model, which is totally reliant on humans known as a specialist for trade execution. The newer model engaged the speed and sophistication of modern microprocessors and computer algorithms, which are faster, better and more accurate than a human specialist to match orders according to a very simple procedure: “Price and Time priority”.

The newer model of exchange is best represented by Electronic Communication Network (ECN) - with the simple mantra of “Who has the best price?” followed by “Who put the best price out first” [23]. These simple mantras has led to a drastic change in the quote execution landscape, putting the latency at which exchanges receive, execute, manage and relay order book information at the forefront of the exchange business model.

The latency difference between execution venues gives rise to several opportunistic strategies directed at Latency Arbitrage, where predatory algorithms execute a trade in one execution venue and offset in another for instant profit. The growing latency lag between the respective liquidity pool has been on the rise, and according to TABB Group, cash trading and market data revenue, which are the most susceptible trade to low-latency competition, have been on a steady increase in the United State equity market and represent over 50% of the total revenue of the US equity market as shown in Figure 2.5 below [23].

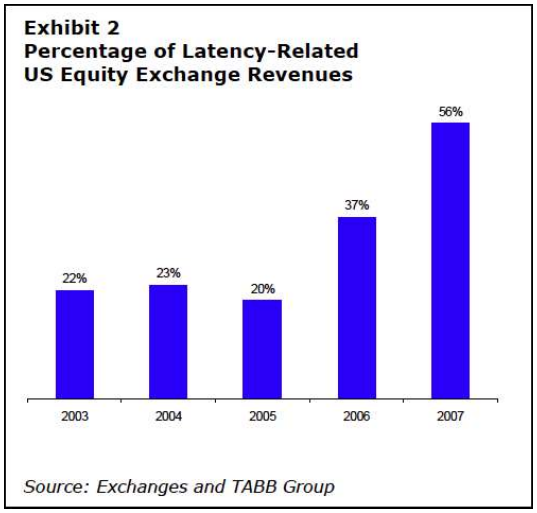


Exhibit 2
Percentage of Latency-Related
US Equity Exchange Revenues

Source: Exchanges and TABB Group

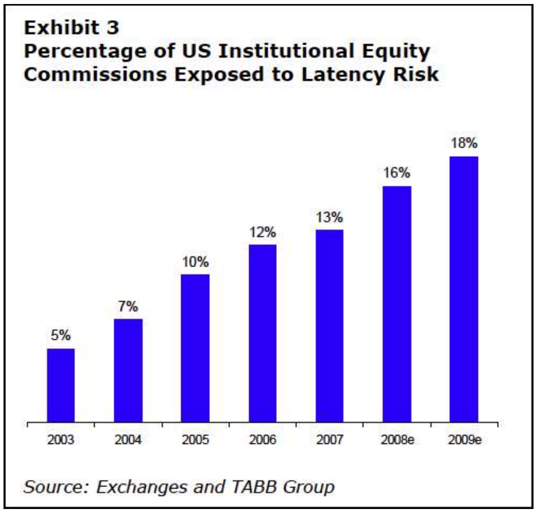
Figure 2.5: Percentage of Latency-Related US Equity Exchange Revenues [24]

Market Players and Latency

The race to zero [25] has become a common term in the financial market, which means the absence of latency ⇒ Latency = 0. The competition for zero latency is more prominent among those market participants that execute market-making strategies (Liquidity providers) to which the majority of HFTs belong. These are HFTs that generate profit by capturing the spread between the supply and demand for a particular security. The agency execution providers are also affected by the latency arm-race. The SOR (Smart Order Router) used to make order route decision relies on real-time market data and thus, a cost is attached to any buy or sell decision if it arrives at a later time than that of a competitor.

The best scenario for this case is the one presented by [26] where the state, “Considering latency from the perspective of a liquidity provider if the presence of observable news in the public domain makes his quote to become stale. It immediately enters a race competition where: 1) Trying to adjust his stale quote and 2) Many others are trying to snap his stale quote. Considering that in a continuous limit order book, messages are processed one at a time in serial and so, even with a cutting-edge speed, one will always lose against many” [26].

The scenario above does not quantify the cost of latency, i.e., the cost implication of being latent, but only presents one of the numerous scenarios where the cost of being latent can lead to a severe adverse effect. The report from TABB Group attempt to put a cost on latency from the perspective of liquidity providers [23], stating that:
“In 2008, 16% of US institutions were exposed to latency risk, totalling $2 Billion in revenue, if a broker’s electronic trading platform is 5 milliseconds behind the competition, it could lose at least 1% of its order-flow which is equivalent to $4 Billion in revenue per millisecond, up to 10ms could result in at least 10% drop in revenue. And finally, if a broker is 100ms slower than the fastest broker, it may as well sell his FIX (Financial Information eXchange) engine and become a floor broker” [23 p.8] (see Figure 2.6).



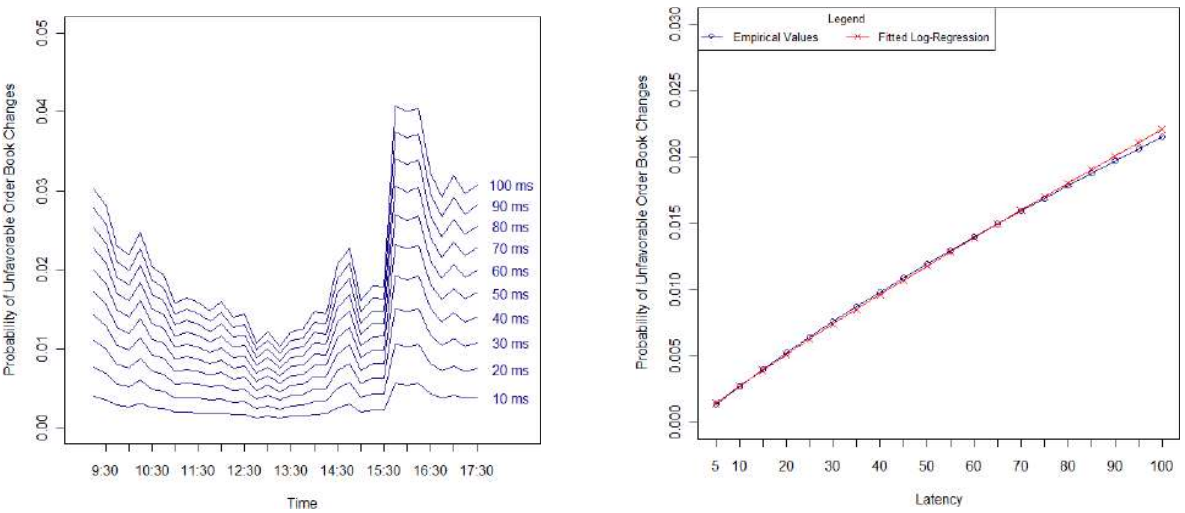
Source: Exchanges and TABB Group

Figure 2.6: Percentage of US Institutional Equity Exposed to Latency Risk

2.5 Impact of HFT Latency On Market Quality

[27] provides a performance measure of the effect of latency in the context of competitive advantage of a trader’s Information Technology (IT) infrastructure over another based on a historical dataset of Deutsche Börse’s electronic trading system “Xetra”. Using trading data from DAX30 instrument traded at Xetra, they used a probability approach to estimate the impact of latency from a trader’s perspective and observed the following:

That the length of latency delay has a considerable impact on the probability of an order book situation changing unfavourably for a submitted order, as shown in Figure 2.7(a).



(a) Day pattern for E.ON and latencies of 10 – 100 ms **(b) Scaling of hit probability due to latency**

Figure 2.7: Latency Impact and Order Book Probability Estimate [27]

The graph in Figure 2.7(a) shows the intra-day patterns for 10 to 100ms for a buy market order in E.ON. It can be observed that in the morning, the probability of the order book alteration is high and decreases continuously due to the fierce competition usually experienced by the exchange during the market opening time - "DAX instruments continuous trading takes place from 9:00 till 13:00 o'clock in the morning and 13:02 till 17:30" [27 p.6].

The probability is observed to reach the minimum just after the midday auction and increase again. The observed striking increase is congruent with the opening time of the US markets, which is at 15: 30 [27]. From this empirical result it can be observed that the cost of latency is higher when competition are high in the market, in the opening of the market; after the midday-auction; during opening hour of other market; or during such event like news. Furthermore, to estimate the cost of being latent, [27] fitted a log-linear regression to the graph of the average increase of probabilities for a buy market order in E.ON as shown in Figure 2.7(b). From the slope of the regression, they deduced a simple rule of thumb. "A 1 % increase in latency leads to a 0.9 % increase in the probability of unfavourable order book changes. Thus reducing latency about 1 ms has a greater effect on the probability" [27 p.8]. Other empirical observations from the historical dataset of Deutsche Börse's electronic trading system "Xetra" as presented by [27] are summarised as follows:

a. Most latency demanders, of which the majority of HFTs belong, concern only the top of the order book i.e most market orders issued either exactly match the best/bid price and volume or are deleted immediately from the order book. This is due to the low latency at the disposal of the HFTs, and thus, it is a more optimal strategy to delete unfulfilled orders from the order book and replace them frequently than leaving them to go stale.
b. Highly capitalised stocks exhibit higher probabilities of encountering order book change than the low capitalised ones from a market participant perspective.
c. The impact of latency is higher on those HFTs whose business model is such that they seek a very low spread, usually, the liquidity providers who execute the market-making strategy by quoting both sides of the order book to profit from the spread, compare to that institution who follows a long-time profit business model - this is usually ATs that are non-HFTs, and of course, human traders or investors.

Furthermore, using the publicly available data from NASDAQ order-level data, identical to those supplied to subscribers and provide real-time information about orders and execution with each entry (order submission, cancellation and execution) time-stamped to milliseconds. [28] observed that the fastest trader have an effective speed of 2 − 3ms, while the slowest speed is observed to be 200ms which are thought to be human traders and not the trades generated by algorithms. They concluded that in the current equity market structure, increased low-latency activities improve liquidity and short-time volatility. The causation explanation for the above observation will be provided later in this research work using the result of the mathematical and simulation model.

It is observed in the reviewed research work that latency impact and cost are estimated based on market data which do not provide an insight into what goes on beyond the exchange. Knowing that in most rapid strategies - those involving low- latency, orders only last a few milliseconds and thus are not executable by the majority of market participants [29]. These failed attempts by relatively slower traders are not captured in the trading data, and thus, analysing latency impact using trading data may be a seemingly-bias approach. But using a simulation approach not only allows for testing different scenarios by varying latency and number of market participants but also allows us to analyse what is happening beyond the exchange by examining and analysing the exchange queue.

3. RESEARCH DESIGN AND METHODOLOGY

Lacking suitable data to empirically study the effect of latency on market quality, a theoretical approach is pursued, which enables the incorporation of causal premises and, specifically, presumptions of how trading behaviour is shaped by environmental conditions [24]. The research uses a Mathematical model underpinned by theoretical observation from available empirical results from research work in the areas of HFT latency and its impact on market quality. The Mathematical model is proposed is validated by providing analysis and explanation consistent with those found in empirical research.

3.1 Mathematical modelling

Models are an abstraction of reality, characterised by assumptions about *variables* (things which change), *parameters* (things which do not change) and *functions* (the relationship between them) [32]. There are different types of mathematical model s, but this research is focused on those applicable to deterministic and stochastic models. Deterministic models have no components that are inherently uncertain, meaning no parameter of the model is characterised by a probability distribution. Unlike stochastic models, a stochastic model will produce many different results depending on the actual value a random variable takes in each realisation [32], which makes it a suitable choice for modelling a financial market where order arrivals and execution are stochastic and not deterministic.

In reality, there exists a large element of compromise in mathematical modelling and modelling in general. The majority of interacting systems in the real world, like the financial market, including its different interacting agents, are very complex and complicated to model in their entirety and thus require a compromise. According to [33], "The best model is the simplest model that still serves its purpose, that is, which is still complex enough to help us understand a system and to solve problems. Seen in terms of a simple model, the complexity of a complex system will no longer obstruct our view, and we will virtually be able to look through the complexity of the system at the heart of things" [33,p.4]

The following steps will be followed in building the mathematical model :

1. Identify the most important part of the real system (the financial system and trader agents) to be included in the model to achieve just the right level of complexity to give a mathematical expression that is worthwhile.
2. Formulate valid assumptions based on the empirical results obtained from research work in the areas of Financial Market modelling and High-Frequency Trading and Latency
3. Identify *variables* (things which change), *parameters* (things which do not change) and *functions* (the

relationship between them)

4. Build a flow diagram and establish a relationship between (3) above based on (2) above
5. Validate the model by testing it against observations from empirical results

4. MARKET QUALITY MODELLING AND ANALYSIS

The mathematical model is developed based on assumptions derived from the results of empirical research as available in the research literature. The resulting model is discussed under the following sub-sections, which are price discovery, liquidity and volatility

4.1 Price Discovery

As observed from the results obtained from the research literature. The contribution of HFT to price discovery is consistent in both theoretical and empirical research. HFT improves price discovery for the faster HFTs but worsens for the slower HFTs leading to adverse selection. The causal explanation for this is provided in the developed Mathematical model .

Price discovery is assumed to be a function of bid-ask activities in the exchange of a financial instrument. With the increase in speed, price discovery becomes more efficient. Naturally, we expect all HFTs to have access to market bid-ask quotes at the same time, but in reality, this is not the case due to the difference in the connection speed between HFT traders and the exchange.

Let the speed between a trader and the exchange = S_t : Where t is the latency (assume to be the time to receive, process, and respond to a quote)

Let the highest speed between a Trader T and the Exchange E = $S_0 \Longrightarrow t = 0$ (latency = 0)

Let the fastest HFT trader in the market be T_{S0}

Let the delay between the fastest HFT trader T_{S0} and an HFT trader T_{St} = $S_0 - S_t = \delta$: Where δ = Latency − Lag

The relationship between Price Discovery Index P, Speed S, Latency t, and Latency-lag δ be:

$$P_t = \exp^{(S_t - \delta)} \qquad (4.1)$$

At the maximum connection speed between the trader T and the Exchange E, $T_{St} = S_0 - S_t = \delta = 0$ (Meaning the Latency Lag $\delta = 0$). Thus, the Price Discovery Index P_0 at latency $t = 0$ can be represented as follows:

$$P_0 = \exp^{(S_0 - 0)} = \exp^{S_0} \qquad (4.2)$$

The graph of P_t and P_0 against S_t and $S_{0,}$ respectively, is shown in Figure 4.1 below:

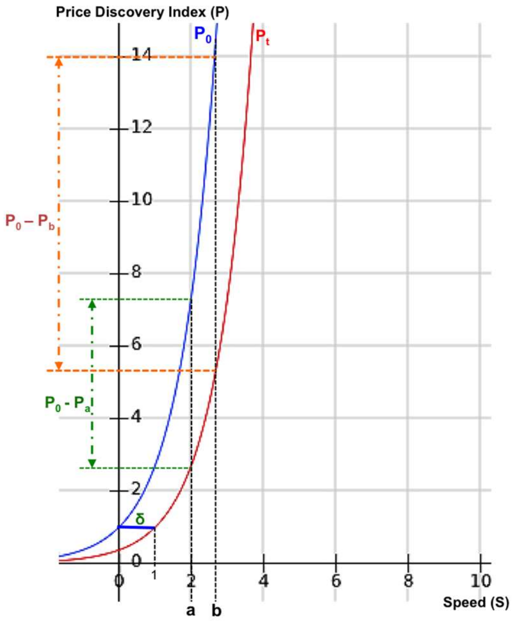


Figure 4.1: Graph of Price Discovery Index (P) Against Speed (S)

From the above graph, as the speed between HFTs and the Exchange increases, the Price Discovery Index P increases exponentially. From the graph, the Latency- lag between the fastest HFT T_0 and a slower HFT T_t = δ = 1. But as the speed S increases from S_a to S_b (from a to b), it can be observed the margin between the price discovery index of the faster HFT trader and the slower HFT trader widened ($P_0 – P_b » P_0 – P_a$). This implies that as the speed between HFTs and the exchange increases, Price discovery improves relatively for the faster HFTs and worsens for the slower HFTs and thus leading to higher transaction cost and adverse selection. This is very logical, for example: If the Latency-lag in terms of time between a faster HFT and a slower HFT is 1ms. An upgrade that reduces the latency at the exchange from 50ms to 10ms, like the upgrade of Deutsche Boerse's trading system on April 23, 2007 [1] will imply an increase in Latency-lag between a fast HFT and slow HFT from 1ms to 5ms.

This means the slower HFTs will always be 5ms behind as against the initial 1ms. This leads to a decrease in price efficiency for the slower HFTs while simultaneously leading to improved price efficiency for the faster HFTs. This agreed with the discovery of [39] that when some HFTs become very fast, it leads to an increase in adverse selection cost for the slower traders and generates negative externalities, which is due to the ability of the faster HFTs to process bid-ask quote information and adjust their trading strategies accordingly before the slower HFTs, leading to adverse selection for the slower HFTs.

Another observation from the Mathematical model is that as the speed of the faster HFTs increases relative to the slower HFTs, the slower HFTs are swept under the carpet, and if intraday trading data are analysed, price discovery will appear to be becoming more efficient, as illustrated in the equation below

From Equation 1 above, $P_t = \exp^{(st-\delta)}$

As $S_t \Rightarrow P_t = \exp^{(st-\delta)} = \exp^{(st)} \longrightarrow \infty$ (4.3)

This implies that the slower HFTs with the Latency-lag δ will no longer be relevant and would, most time withdraw from the market as observed by [39] that in equilibrium, smaller institutions (slower HFTs) are less informed than their large counterpart (faster HFTs) and exit the market when HFTs become more prevalent to avoid adverse selection cost.

Under the assumption of predominant speed-based competition, we can conclude that the relationship between Price Discovery Index P , Speed S , Latency t and Latency-lag δ can be expressed as an exponential function

$f_{(S\ )} = \exp^{(st-\delta)} = P_t$ (4.4)

Where δ is the latency-lag between the fastest HFT and a slower HFT under consideration.

4.2 Liquidity

The assumption of predominant speed-based competition implies all HFTs will want to buy or sell a financial instrument at almost the same time. A good example is the release of a negative news about a stock, where we would expect all HFTs to initiate a sell order with no HFT initiating a buy order at the other end of the market, thus liquidity should disappear instantaneously. But this does not happen due to disparity in the different HFT's speed. Instead, the slower HFTs will initially provide liquidity due to the delay in receiving the news, but revert their position as soon as such HFTs become aware of the news. This continues until the slowest HFT receive the news. At this point liquidity will momentarily disappear unnoticeably because the faster HFT will immediately start liquidating their position by buying back from the slower HFTs at a lower price than they initially sold, profiting from the bid-ask spread. While the slower HFTs will be selling to the faster HFT at a price lower than they bought, thereby recording a loss and leading to adverse selection.

Let L be the Liquidity Index
Let the latency between any HFT and the Exchange be = t
Let the latency for the fastest HFT in the market be = $\delta_t = \delta_0 = 0$
Let the latency for the slowest HFT in the market be = $\delta_t = \delta_\infty = 1$
Let the Latency-lag between the fastest HFT and any slower HFT be = $\delta_0 - \delta_t = \theta$
Let the relationship between L, δ_0, δ_t be expressed as follows:

$L = -\log[1 - (\delta_0 - \delta_t)] = -\log(1 - \theta)$: Where, $\theta = \delta_0 - \delta_t$ (4.5)

The graph of L against θ is shown in Figure 4.2 below:

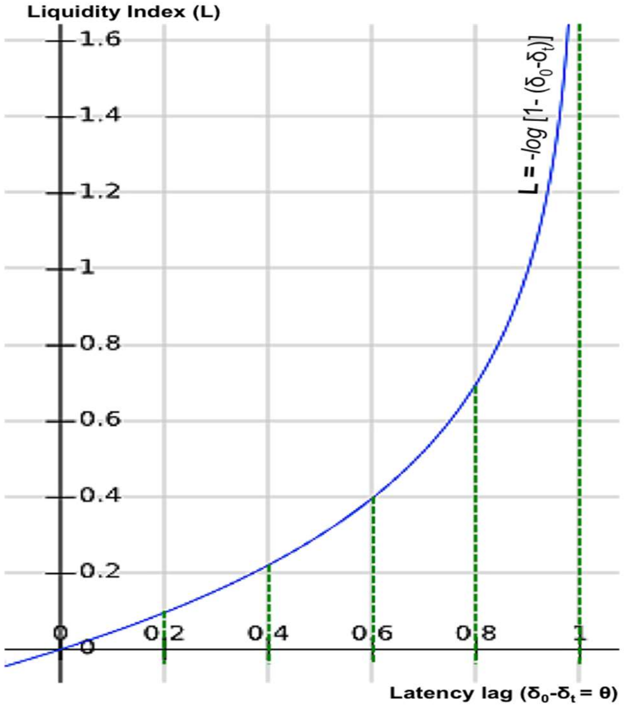


Figure 4.2: Graph of Liquidity Index (L) Against Latency Lag (θ)

As shown above, Liquidity for a particular financial instrument reduces as the latency-lag (θ) reduces (The difference between the time it takes a slower HFT to receive information from the market minus that of the faster HFT), meaning more HFTs are getting informed. For example, this could be a case of a negative news, a large market order by a big AT firm or as a result of a large market movement in a particular direction. At the point where Latency lag (θ) = 0, Liquidity momentarily = 0 as shown in Figure 4.2 below. This signifies a point of maximum information circulation. At this point, the faster HFTs starts liquidating their position - since they will be the first to be aware of the illiquidity, and start to stabilise the price and thus making profit from the bid-ask spread while generating adverse selection cost for the slower HFTs.

[40] observation when measuring the contribution of HFT and non-HFT liquidity supply and liquidity demand to price change com- ponents and found out that HFTs trades in opposite direction to the market price movement can be explained from the above model. HFTs, in general, consume liquidity before providing it. This makes it seems HFTs are actually stabilising price, but in reality, the faster HFTs use their speed advantage to momentarily take away liquidity and quickly liquidate their position, making a profit from the bid-ask spread, which shall be proved experimentally by simulation modelling in the next chapter. [40] also observed that HFTs consume liquidity when the bid-ask spread is low and provide liquidity when the bid-ask spread is wide. As shown in the graph of the Mathematical model above, when the latency-lag θ is maximum, the bid-ask spread is at the minimum. But as soon as the faster HFTs start reacting to market information, for example, by selling - in the case of negative news, they begin to take away liquidity and introduce volatility. At the point where liquidity momentarily = 0, bid-ask spread = Maximum, and thus the faster HFTs start providing liquidity. This is a better causal explanation for [40] observation - be proved experimentally by simulation modelling in the next chapter.

The Model also provides an explanation for the flash crash that occurred on May 6, 2010. The Mathematical model shows that the faster HFTs wait until the momentary time where liquidity equals zero before liquidating their position. But in the presence of a large order, as witnessed in the flash crash [8], HFTs will continue to consume liquidity as long as it is available, driving down the price and at the same time, increasing volatility until a point of illiquidity is reached before liquidating their position. The presence of a very large order will make HFTs not liquidate their position, and this may lead to a crash if no mechanism is in place to manually pause trade on such stock.

The correlation of the result of the Mathematical model with empirical results shows that the assumption of HFT predominant speed based-competition holds and that Latency positively impacts liquidity but at the same time generates negative externalities for the slower HFTs. Thus, the relationship between Liquidity Index L and the Latency lag θ can be expressed as follows:

$f_{(\delta 0-\delta t)} = -\log[1-(\delta_0-\delta_t)] = L$ (4.6)

Since $\delta - \delta_t = \theta$

$f_{(\theta)} = -\log(1-\theta) = L$ (4.7)

Furthermore:

$\theta \Rightarrow f_{(\theta)} = -\log(1-0) = 0 \Rightarrow L = 0$ (4.8)
$\rightarrow 0$

This shows that if the latency lag θ equals 0, meaning that all HFTs have equal latency, and thus liquidity will be equal to Zero. Even though it is not possible for latency to be equal, but if the slower HFTs withdraw from market as HFTs intensity increases leaving the faster HFTs [8], a condition close to this may occur. To overcome this, exchanges provides rebate to liquidity providers to encourage liquidity and also, some HFTs are registered as liquidity providers to avoid this condition.

Also:

$\theta \Rightarrow f_{(\theta)} = -\log(1-1) \Rightarrow L =$ ”Undefined” (4.9)
$\rightarrow 1$

This means that, at the point of maximum latency lag θ = 1 there is virtually no activity in the market and thus Liquidity L does not exist. This implies a crash as witnessed on 6th May, 2010.

4.3 Volatility

As shown from the empirical results from the literature, [22] studied the effect of HFT intensity on market liquidity, short-term volatility, and price discovery between 2001 and 2011 in 42 equity markets around the world and found out that, on average - HFT improves liquidity and price discovery but increases volatility. [39] also analysed trading equilibrium for a given level of HFT and discovered that when some HFT becomes very fast, it leads to in- creased volatility and thus increased adverse selection cost for the slower traders and generates negative externalities. This shows that when HFTs' intensity in the market increases, this lead to volatility. This condition of HFT intensity leading to an increase in volatility is only possible if most of the HFTs are trading on the same strategy, meaning buying or selling at the same time. This creates a condition of price imbalance and results in volatility. This agreed with the assumption of predominant speed-based competition. Volatility can be modelled under this assumption as follows.
Let V be the Volatility Index

Let the latency between any HFT and the market be = t
Let the latency for the fastest HFT in the market be = $\delta_t = \delta_0 = 0$

Let the latency for the slowest HFT in the market be = $\delta_t = \delta_\infty = 1$

Let the latency difference between the fastest HFT and any slower HFT be =

$$\delta_0 - \delta_t = \theta$$

Let the relationship between V, δ_0, δ_t be expressed as follows:

$$V = -\log(\delta_0 - \delta_t) = -\log(\theta) : \text{Where, } \theta = \delta_0 - \delta_t \qquad (4.10)$$

The graph of V against θ is shown in Figure 4.3 below:

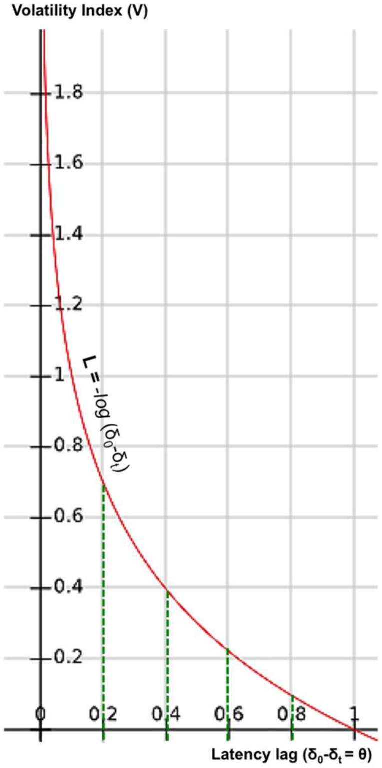


Figure 4.3: Graph of Volatility Index (V) Against Latency Lag (θ)

It can be observed from the graph that, as the Latency lag θ decreases, volatility increases. The Latency lag θ is the latency difference between the fastest HFT trader and a slower Trader, as this difference reduces, meaning that the slower HFTs are becoming more informative and thus execute a trade in the same direction as the faster HFTs and thus leading to increased volatility. Take, for example, The arrival of a large market bid order to buy a financial instrument. The faster

HFTs will be the first to be aware, assuming that the firm placing such an order may have access to private information and start issuing a market buy order. The other HFTs will be providing liquidity until they become aware of the trends and start to liquidate their acquired position by buying back which leads to increased volatility on the particular security. This agrees with the result of Bias and Foucault [9] that when some HFT becomes very fast, it leads to increased volatility and thus increased adverse selection cost for the slower traders and generates negative externalities.

At the point of highest volatility, that is, when all traders are now fully aware of the market trend (when latency lag θ = 0). The slower HFTs withdraw from the market to avoid adverse selection costs, at this point, the faster HFTs start liquidating their position by selling. This agreed with the result of [40]. They observed that HFTs consume liquidity when the bid-ask spread is low (low volatility) and provide liquidity when the bid-ask spread is wide (high volatility).

From Equation 4.13 above, it can be observed that:

$$\theta_{\to 0} \Rightarrow f_{(\theta)} = -\log(0) \Rightarrow V = \text{"Undefined"} \qquad (4.11)$$

This implies that, at the point of highest volatility - when θ ≈ 0, if the faster HFTs do not liquidate their position after the withdrawal of the slower HFTs, the market could be heading for a crash. There is usually a mechanism in place to checkmate this in the security market.
Also:

$$\theta_{\to 1} \Rightarrow f_{(\theta)} = -\log(1) \Rightarrow V = 0 \qquad (4.12)$$

This means that, at the point of highest latency, volatility $V = 0$. This implies a trade-off between low latency $\theta \approx 0$ and high latency $\theta \approx 1$, and thus the optimum latency should lie in between. That is, for Optimum Volatility V_0 and Optimum Latency θ_0

$$V_0 = f_{(\theta 0)} : (0 < \theta_0 < 1) \qquad (4.13)$$

There have been numerous propositions, both from the academic and the financial sectors about what the optimum latency should be. [26] proposed a discrete-time frequent batch as a replacement for the continuous-time limit order book where the market and limit Bid/Ask Orders are processed in batches by the Exchanges based on a fixed-time interval. The fixed-time interval at which an Exchange processes an order is the respective "induced latency" of the exchange as proposed [26]. Thus, the optimum speed, which is the minimum speed an HFT must possess to be able to compete without negative externalities as a result of being latent, will be the fixed-time interval at which the Exchange processes Bid/Ask Orders.

Let δ_i = Induced-latency (Latency at which the exchanges processes Orders based on the proposed discrete-time frequent batch by Budish et al [26].

Let the latency of an HFT under consideration be = δ_{HFT}

Let the Optimum-latency based on the proposed discrete-time batch be the latency at which $\delta_{HFT} \geq \delta_i$

Thus:

$$\delta \text{Optimum} \Rightarrow \delta \text{HFT} \geq \delta \text{i} \qquad (4.14)$$

The above condition will allow for a bounded latency competition where HFTs will only strive to achieve the optimum latency to remain competitive and not the unbounded latency competition as we have today in the financial market where the fastest wins it all. The term "fastest" is a very relative term in the context of HFT, since yesterday's fastest speed could mean today's slowest. The only limitation of HFT latency in today's competition is the natural barrier poised by nature to achieving Zero latency and the limitation of innovation in the areas of data transmission and communication.

The value of the optimum latency based on the proposed Mathematical model and Budish et al [26] could best be determined empirically, which is beyond the scope of this research work. However, we are able to show theoretically that the optimum latency should be between the highest and lowest latency obtainable in the market where Volatility is optimum, as shown below.

$$V_0 = f_{(\theta 0)} \Rightarrow (0 < \theta_0 < 1) \qquad (4.15)$$

Where $\theta_0 \Rightarrow$ Optimum Latency

And based on Budish et al proposal [26]

$$\delta \text{Optimum} \Rightarrow \delta \text{HFT} \geq \delta \text{i} \qquad (4.16)$$

Where $\delta_{Optimum}$ = Optimum Latency

δ_{HFT} = Latency of HFT under consideration and

δ_i = Induced Exchange latency based on the proposed discrete-time frequent batch by Budish [26].

5. CONCLUSIONS

This research work is focused on the analysis and evaluation of the impact of HFT latency on market quality which are liquidity, price discovery and volatility. To achieve this, the HFT latency in relation to the market microstructure was analysed from both the perspective of the exchange and that of market participants, which are the High-Frequency Traders (HFTs), Algorithmic Traders (ATs), the Investors and human traders. Also, mathematical modelling theory and application were reviewed, including its application in financial intelligence. Furthermore, a mathematical model was developed based on the empirical result from the research literature. Certain assumptions were made to theorise the impact of latency on market quality.

The output of the research serves to provide the foundation for developing a simulation model for market quality underpinned by the research output in HFTs and non-HFTs activity in the liquidity market. In such a simulation model underpinned by the proposed mathematical model, different market parameters can be varied, such as latency and market participants' behaviour, to observe the impact of these parameters on market quality, such as price discovery, liquidity, and volatility.

REFERENCES

[1] Riordan, R., & Storkenmaier, A. (2012). Latency, liquidity and price discovery. Journal of Financial Markets, 15, 416–437.

[2] CFA Magazine: The Impact of High-Frequency Trading on Markets [Online]. Available at: https://www.rbccm.com/globalequity/file-569694.pdf [Accessed on: 20 June 2015].

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

[4] Gomber,P., Arndt,B.,Lutat,M., Uhle,T. (2011) High Frequency Trading [Online]. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1858626 [Accessed on: 1 June 2015].

[5] Brogaard, J. A. (2010). High frequency trading and its impact on market quality. Northwestern University Kellogg School of Management Working Paper [Online]. Available at: http://heartland.org/sites/default/files/htf.pdf [Accessed on: 20 June 2015].

[6] Jones, C. M. (2013). What do we know about high-frequency trading? Charles
M. Jones* Columbia Business School Version 3.4: March 20, 2013. Columbia Business School.

[7] Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does Algorithmic Trading Increase Liquidity? Journal of Finance, 66(1), 1–33.

[8] Kirilenko, A., Kyle, A., & Samadi, M. (2010). The flash crash: The impact of high-frequency trading on an electronic market [Online]. Available at: http://www.ftm.nl/wpcontent/ uploads/content/files/Onderzoek%20Flash%20Crash.pdf [Accessed on: 20 June 2015].

[9] Ye, M. (2012). The Externalities of High Frequency Trading, 1–48 [Online]. Available at: http://www.sec.gov/divisions/riskfin/seminar/ye031513.pdf [Accessed on: 20 June 2015]

[10] Zhang, X. F. (2010). The Effect of High-Frequency Trading on Stock Volatility and Price Discovery, 1–53. [Online]. Available at: http://mitsloan.mit.edu/groups/template/pdf/Zhang.pdf [Accessed on: 20 June 2015]

[11] Menkveld, A. J. (2013). High-frequency trading and the new market makers. Journal of Financial Markets, 16(4), 712–740.

[12] Moallemi, C. C., & Sağlam, M. (2013). OR Forum—The Cost of Latency in High Frequency Trading. Operations Research, 61(5), 1070–1086. doi:10.1287/opre.2013.1165

[13] Viraf,W.(2008). TheValue of a Millisecond: Findingthe Optimal speed of a Trading Infrastructure [Online]. Available at: http://www.tabbgroup.com/PublicationDetail.aspx?PublicationID=346 [Accessed on: 20 June 2015]

[14] Ende, B., Uhle, T., & Weber, M. C. (2011). The Impact of a Millisecond
: Measuring Latency Effects in Securities Trading. Proceedings of the 10th International Conference on Wirtschaftsinformatik (WI), (February), 27–37.

[15] Dhahahjay, K.D. & Shyam, K. (n.d.). Allocative Efficiency of Markets With Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality. The Journal of Political Economy, 101(1), 119–137.

[16] Farmer, J. D., Patelli, P., & Zovko, I. I. (2005). The predictive power of zero intelligence in financial markets. Proceedings of the National Academy of Sciences of the United States of America, 102, 2254–2259.

[17] Brogaard, J. (2010). High Frequency Trading and its Impact on Market Quality, 5th Annual Conference on Empirical Legal Studies Paper. http://ssrn.com/ paper=1641387.

[18] Chaboud, A., Erik, H., Clara, V., & Ben, C. (2009). Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market. [Online]. Available at: http://ssrn.com/ paper=1501135 [Accessed on: 21 June 2015]

[19] Gomber, P., Arndt, B., Lutat, M., & Uhle, T. (2011) High-Frequency Trading [Online]. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1858626 [Accessed on: 12 July 2015]

[20] Hendershott, T., & Riordan, T. (2011), "High frequency trading and price discovery" working paper, UC Berkeley.

[21] Biais, B., & Moinas, S. (2011). Equilibrium High Frequency Trading [Online]. Available at: http://www.lse.ac.uk/fmg/events/conferences/pastconferences/ 2012/PWC-Conference_7-8June/Papers-andslides/ Bruno_Biais_paper.pdf [Accessed on: 22 July 2015].

[22] Boehmer, E., Fong, K., & Wu, J. (2012). International evidence on algorithmic trading. Available at SSRN 2022034, 2012, 49.

[23] Viraf, W. (2008). The Value of a Millisecond: Finding the Optimal speed of a Trading Infrastructure, (April). Retrieved from http://www.tabbgroup.com/PublicationDetail.aspx?PublicationID=346

[24] Wah, E., & Wellman, M. (2013). Latency arbitrage, market fragmentation, and efficiency: a two-market model. Proceedings of the Fourteenth ACM Conference on Electronic Commerce, 1(212), 855–872. Retrieved from http://dl.acm.org/citation.cfm?id=2482577

[25] Barr, D., Benos, E., Braun-munzinger, K., Butterworth, E., Chichkanov, P., Cornelius, M., Meeks, R. (2012). Financial arms races, (April). Based on a speech delivered at the Institute for New Economic Thinking, Berlin 14 April 2012. [Online]. Available at: http://www.bankofengland.co.uk/publications/Documents/speeches/2012 /speech565.pdf [Accessed on: 20 June 2015]

[26] Budish, E., Cramton, P., & Shim, J. (2014). A Market Design Approach to the HFT Debate : The Case for Frequent Batch Auctions. [Online]. Available at: http://faculty.chicagobooth.edu/eric.budish/research/HFT-FrequentBatchAuctions.pdf [Accessed on: 20 June 2015]

[27] Ende, B., Uhle, T., & Weber, M. C. (2011). The Impact of a Millisecond : Measuring Latency Effects in Securities Trading. Proceedings of the 10th International Conference on Wirtschaftsinformatik (WI), (February).

[28] Hasbrouck, J., & Saar, G. (2013). Low-latency trading. Journal of Financial Markets, 16(4), 646–679.

[29] Goldstein, M. A., Kumar, P., & Graves, F. C. (2014). Computerized and High-Frequency Trading, 49(2), 1–35.

[30] Gode, D. K. and Sunder, S. (1993). Allocative efficiency of markets with zero- intelligence traders: Market as a partial substitute for individual rationality. Journal of Political Economy 101, 1, 119–137.

[31] Farmer, J. D., Patelli, P., & Zovko, I. I. (2005). The predictive power of zero intelligence in financial markets. Proceedings of the National Academy of Sciences of the United States of America, 102, 2254–2259.

[32] Bokil, V. A. (2009). Introduction to Mathematical modelling. [Online]. Available at: http://www.mesacc.edu/~davvu4111/IntroToModel.pdf [Accessed on: 20 June 2015].

[33] Dym, C. L. (1980). Principles of Mathematical modelling. American Journal of Physics, 48(11), 994.

[34] Morgan, C. B., Banks, J., & Carson, J. S. (1984). Discrete-Event System Simulation. Technometrics, 26(2), 195.

[35] Karnon, J., Stahl, J., Brennan, A., Caro, J. J., Mar, J., & Möller, J. (2012). Modelling using Discrete Event Simulation: A Report of the ISPOR-SMDM Modelling Good Research Practices Task Force-4 Background to The Task Force. Value in Health, 15(6), 821–827.

[36] Mathworks SimEvents User's Guide for R2015a (2015). [Online]. Available at: http://cn.mathworks.com/help/pdf_doc/simevents/simevents_ug.pdf [Accessed on: 20 June 2015]

[37] Zheng, H., Son, Y.-J., Chiu, Y.-C., Head, L., Feng, Y., Xo, H., . . . Hickman, M. (2013). A Primer for Agent-Based Simulation and Modelling in Transportation Applications.

[38] Maidstone, R. (2012). Discrete Event Simulation, System Dynamics and Agent Based Simulation: Discussion and Comparison. System, 1–6.

[39] Biais, B., & Moinas, S. (2011). Equilibrium High Frequency Trading [Online]. Available at: http://www.lse.ac.uk/fmg/events/conferences/pastconferences/ 2012/PWC-Conference_7-8June/Papers-andslides/ Bruno_Biais_paper.pdf [Accessed on: 22 April 2015]

[40] Hendershott, T., & and Riordan, T. (2011), "High frequency trading and price discovery" working paper, UC Berkeley

[41]Mathworks Simulink Intro R2015a (2015). [Online]. Available at: http://uk.mathworks.com/products/simulink/ [Accessed on: 20 June 2015]

[42] Davis, J. P., Bingham, C. B., & Eisenhardt, K. M. (2007). Developing Theory Through Simulation Methods. Academy of Management Review, 32(2), 480–499.

[43] Bonabeau, E. (2009). Decisions 2.0 : The Power of Collective Intelligence. MIT Sloan Management Review, 50(50211), 45–52.