Accepted author manuscript. To appear in Proceedings of 2018 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr-2018), Nov 18-21, Bengaluru, India

Agent-Based Model Exploration of Latency Arbitrage in Fragmented Financial Markets

Matthew Duffin* and John Cartlidge†
Department of Computer Science
University of Bristol
Bristol, UK
Email: *md14816.2014@my.bristol.ac.uk, †john.cartlidge@bristol.ac.uk

***Abstract*—Computerisation of the financial markets has precipitated an arms-race for ever-faster trading. In combination, regulatory reform to encourage competition has resulted in market fragmentation, such that a single financial instrument can now be traded across multiple venues. This has led to the proliferation of high-frequency trading (HFT), and the ability to engage in latency arbitrage (taking advantage of accessing and acting upon price information before it is received by others). The impact of HFT and the consequences of latency arbitrage is a contentious issue. In 2013, Wah and Wellman used an agent-based model to study latency arbitrage in a fragmented market. They showed: (a) market efficiency is negatively affected by the actions of a latency arbitrageur; and (b) introducing a discrete-time call auction (DCA) eliminates latency arbitrage opportunities and improves efficiency. Here, we explore and extend Wah and Wellman's model, and demonstrate that results are sensitive to the bid-shading parameter used for zero-intelligence (ZIC) trading agents. To overcome this, we introduce the more realistic, minimally intelligent trading algorithm, ZIP. Using ZIP, we reach contrary conclusions: (a) fragmented markets *benefit* from latency arbitrage; and (b) DCAs *do not* improve efficiency. We present these results as evidence that the debate on latency arbitrage in financial markets is far from definitively settled, and suggest that ABM simulation—a form of decentralised collective computational intelligence—is a productive method for understanding and engineering financial systems.**

***Index Terms*—Agent-Based Modelling, Continuous Double Auction, Discrete-time Call Auction, Latency, Arbitrage, Financial Markets, Fragmentation, ZIP, ZIC, High-Frequency Trading**

I. Introduction

Computerisation of the financial markets and the competition for speed has resulted in the rise of high-frequency trading (HFT)—the use of algorithms to trade at superhuman speeds with no human intervention—such that positions are often held for a fraction of a second. As markets have become faster, they have also become more fragmented. Regulatory reforms to encourage competition in the markets, such as RegNMS in the USA and MiFID in Europe, have resulted in a proliferation of new exchange venues. Therefore, where there would once have been a central market venue bringing traders together, modern markets contain multiple venues across which a single financial instrument can be traded.

Market fragmentation leads to difficulties in traders accessing up-to-date price information across the market. To address this issue, RegNMS requires exchanges in the USA to provide pricing information to a central entity called the Securities Information Processor (SIP), which in turn publishes a snapshot of the best prices currently available across all markets—the National Best Bid and Offer (NBBO).

However, since venues are geographically dislocated, there is necessarily a delay in the NBBO. Therefore, if a faster third party HFT is able to obtain the same pricing information and calculate the NBBO before SIP's official NBBO is published, it may be possible for the HFT to identify price disparities before they become public. Utilising this information, HFT algorithms may be able to "jump ahead" of competitors to secure a deal. We consider this form of trading as *latency arbitrage*. Arbitrage in general is the practice of taking advantage of disparities in prices between different markets in order to secure a profit at zero (or minimal) risk; for example, by simultaneously buying at $99 in one market and selling at $100 in another, for a guaranteed risk-free profit of $1. Latency arbitrage exploits arbitrage opportunities that exist over short time-scales due to delays in dissemination of information.

Debate rages strong about latency arbitrage, with some considering it a predatory practice and evidence of a "rigged" market [1], while others (most often HFT practitioners) proclaim the moralistic virtues of HFT as a beneficial liquidity-providing service [2]. This polarisation of opinion is indicative of a highly secretive industry in which "everyone knows that loose lips get pink slips" [3]. The details of algorithmic trading is considered an existential trade secret to most HFT firms, leading to a paucity of information for regulators and researchers. As a result, empirical estimates of latency arbitrage profits using historical trading data vary considerably. In 2012, aggregated annual profits from latency arbitrage was estimated at $21bn [4]. Potential profits from latency arbitrage opportunities in S&P500 symbols during 2014 was estimated at $3.03bn [5]. HFT profits (all strategies, not just latency arbitrage) for 2013 across all US share trading was estimated at $1.25bn [2]. While, in 2017, it was argued that latency arbitrage is not a meaningful source of HFT profits [6].

To shed light on this issue and to gain a better theoretical understanding at the system level, here we use an agent-based model (ABM) to simulate the effects of latency arbitrage in fragmented markets. Building upon the ABM introduced by Wah and Wellman [7], the model contains a population of

†Contact author, John Cartlidge, is sponsored by Thomson Reuters.

traders, multiple continuous double auction (CDA) exchange venues, a SIP entity that processes market prices and produces an NBBO (with a delay, controlled by a simulation parameter, δ), and a latency arbitrageur HFT that has instant access to all market prices. We initially validate the model implementation by replicating Wah and Wellman's results (Section V), before rigorously exploring and extending the model in Section VI. We show that the purely stochastic, zero-intelligence (ZIC) agents [8] used in the original model are sensitive to a bid-shading parameter, R, which controls the markup (or profit margin) that agents attempt to secure. To counter this, we introduce ZIP agents—adaptive trading agents that update profit margin based on prevailing market conditions [9]. By using a simple machine learning rule to maximise profits, ZIP agents have a minimal intelligence, unlike purely stochastic zero-intelligence agents. Populations of ZIP agents can therefore be considered as a decentralised collective computational intelligence. ZIP agents have previously been shown to be a better model of human trading; demonstrating more realistic market dynamics [9], and capable of outperforming human traders [10]. We show that—counter to the results when using ZIC agents—fragmented markets populated with ZIP agents *benefit* from latency arbitrage; while moving from a continuous double auction (CDA; the predominant auction mechanism used in financial markets) to a discrete-time call auction (DCA; intended to reduce latency arbitrage opportunities) *reduces* overall market efficiency. These results have the potential to impact understanding and regulation of modern financial markets, while demonstrating a method for engineering better financial systems.

II. Related Work

Here, we briefly summarise related studies of HFT and latency arbitrage. We begin with empirical studies using real-world trading data, move on to theoretical models of latency arbitrage, and end with models of market fragmentation.

A. Empirical studies of HFT

By determining trading strategies using NASDAQ order data to construct linked sequences of messages, Hasbrouck and Saar [11] study the effects of low-latency activity—strategies responding to market events in the millisecond environment—on dimensions of market quality. Their analysis suggests low-latency activity *improves* traditional market quality measures.

Kirilenko et al. [12] study audit-trail data from the 2010 *flash crash* to discover how HFT traded on that day. They conclude that while HFT were not responsible for the flash crash, their activity *exacerbated* market volatility.

Baron et al. [13] examine trading in one E-mini S&P 500 futures contract over one month. They find that high frequency traders are highly profitable, and extract over $29 million of profit. This mostly comes at the expense of opportunistic investors, but also harms institutional and retail investors.

B. Models of latency arbitrage

Cohen and Szpruch [14] build a mathematical model of a single market containing two investor types: fast, and slow. They demonstrate that if the faster trader is able to predict the trades that the slower trader will make, the fast trader can realise a risk-free profit. They also show that the response slower traders might take when knowing faster traders are present, can *reduce* market efficiency. Finally, the introduction of a transaction tax is shown to prevent the fast trading strategy.

Hanson [15] makes use of agent-based modelling to model a continuous double auction market with multiple high frequency traders present. Liquidity, efficiency, and surplus increase when the number of traders employing the HFT strategy is increased. However, volatility also increased and the profits of the HFT traders were found to possibly come at the expense of long-term investors.

Jarrow and Protter [16] use a mathematical model to show that HFT can unknowingly cause mispricings (deviations from the fundamental value) by collective and independent actions co-ordinated by the observation of common signals.

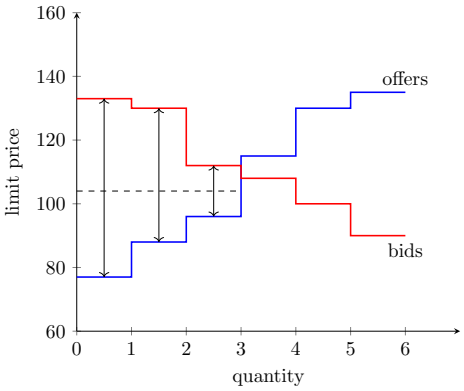
C. Modelling market fragmentation

Mendelson [17] presents a theoretical analysis of the effects of consolidation versus fragmentation in periodic call markets; finding that fragmentation reduces trade quantity, increases volatility, and reduces overall surplus. However, the study does not consider arbitrage between the fragmented markets.

Relatively few prior studies attempt to model latency between fragmented markets and the arbitrage opportunities that result. Ding et al. [18] use proprietary data feeds from stock exchanges, as well as public NBBO data feeds, to provide the first public evidence that direct access to exchanges and fast calculation of the NBBO can be used to generate profits. Open questions are raised about how well current regulation meets its goals; including whether or not there is sufficient incentive for public data providers to reduce NBBO latency.

Discrete-time call auctions (DCA) have previously been proposed as an alternative to continuous double auctions (CDA) due to their ability, by design, to remove arbitrage opportunities. Wah and Wellman [7], [19] use an agent-based approach and a two-market model to study the effects of the presence of a latency arbitrageur in fragmented markets; demonstrating that a DCA provides efficiency gains. Wah, Hurd and Wellman [20] consider market choice: if DCA markets were available, would anybody use them? To model this, they give fast and slow agents access to both a CDA market and a frequent DCA. At equilibrium, they find that the welfare of the slow agents is generally better in the call market. They also show that the slow traders seek the refuge of the frequent DCA when preyed upon by faster traders.

The relative speed of agents has also been shown to affect market efficiency in experimental human versus agent market experiments. The presence of fast traders has been shown to reduce market efficiency [21], and can lead to endogenous fragmentation within a single market, such that fast (slow) traders are more likely to execute with fast (slow) traders [22].



(a) Graph of supply and demand.

Bids	Offers
133	77
130	88
112	96
108	**115**
100	**130**
90	**135**

(b) LOB.

Fig. 1: Clearing operation in a DCA: (a) vertical arrows indicate orders matched at uniform clearing price, 104 (horizontal dashed line); (b) post-clearing, bold orders remain in the LOB.

In this paper, we explore and extend the agent model of market fragmentation used by Wah and Wellman [7]. Next, we introduce the necessary technical background.

III. Technical Background

A. Continuous Double Auction (CDA)

The continuous double auction (CDA) is the predominant auction mechanism used in today's financial markets. The CDA enables buyers and sellers to submit *bids* (orders to buy) and *offers* (orders to sell) at any time. Limit orders specify a price beyond which the order will not execute. For a buyer, the bid limit is the maximum price a trader is willing to buy; for a seller, the offer limit is the minimum price a trader is willing to sell. Orders arrive at the exchange and—if not immediately executed—are stored in the *limit order book* (LOB). The LOB stores bids ordered by price in descending order, and offers ordered by price in ascending order. Thus, the *top* of the LOB displays the best bid price (p_{bb}, the highest priced buy order) and best offer price (p_{bo}, the lowest priced sell order); commonly referred to as the best bid and offer (BBO). When a new bid to buy (with limit price p_{nb}) enters the exchange, it will immediately match with the current best offer to sell if $p_{nb} \geq p_{bo}$. A trade will execute at price p_{bo}; the price of the original resting offer, not the price of the new incoming order. Likewise, when a new offer to sell (with limit price p_{no}) arrives, it will match with the current best bid if $p_{no} \leq p_{bb}$; triggering a trade at price p_{bb} (the price of the original resting bid). When a new order does not immediately execute, it is stored in the LOB (sorted by price). The exchange publishes all trades and the current state of the LOB to all traders.

B. Discrete-Time Call Auction (DCA)

In contrast with the immediate execution of the CDA, a discrete-time call auction (DCA) batches together groups of orders that arrive during a discrete time interval. At the end of each time interval, a clearing operation is performed to match orders and execute trades at a uniform price. Fig. 1 demonstrates how clearing operates. The bid with the highest limit price, BID_{MAX}, is matched with the offer with the lowest limit price, $OFFER_{\text{MIN}}$. Next, the second highest bid is matched with the second lowest offer. This *uncrossing* process continues until orders can no longer be matched; i.e., when $BID_{\text{MAX}} < OFFER_{\text{MIN}}$. A uniform transaction price for all matched trades is calculated as the midpoint between the lowest matched bid and the highest matched offer. In Fig. 1, uniform transaction price is the midpoint between 112 and 96 ($= 104$; as indicated by the dotted line). The visible BBO immediately after the clear is (108, 115). The exchange only publishes information about the contents of the LOB immediately after clearing, thus making DCA a sealed-bid auction. By clearing at regular intervals, the DCA eliminates latency arbitrage opportunities as potential latency arbitrageurs are unable to exploit informational advantages within the *clearing interval*.

C. Zero-Intelligence Constrained (ZIC) Traders

Introduced by Gode and Sunder [8] to evaluate efficiency and equilibration in CDA markets, Zero-Intelligence Constrained (ZIC) agents are simple traders that submit bid and offer prices sampled from a uniform distribution. ZIC traders are *constrained* not to enter into loss-making trades, therefore bid and offer prices for agent i are bounded on one side of the distribution by the agent's private value, PV_i. In this study, the other side of the distribution is bounded by a *bid shading* parameter, R (alternatively referred to as *markup*, or *margin*). For buyer, i, with private value, PV_i, bid price $b_i \in U[PV_i - R, PV_i]$; for seller, i, offer price $s_i \in U[PV_i, PV_i + R]$. Despite the deliberately simplistic design of ZIC, populations of ZIC traders have been shown to reach allocative efficiency close to 100% [8]; however, in markets where demand and supply is asymmetric, performance can be significantly lower [9].

D. Zero-Intelligence Plus (ZIP) Traders

Zero-Intelligence Plus (ZIP) agents were developed to overcome the provable shortcomings of ZIC agents [9]. ZIP agents are profit-driven traders that adapt using a simple learning mechanism; adjusting profit margin, μ, to generate a competitive order price, p, based on the current price, q, of the most recent bid/offer in the market.

At time t, a ZIP trader, i, with limit price (private value), λ_i, uses margin, μ_i, to calculate new order price, p_i, as:

$$p_i(t) = \lambda_i(1 + \mu_i(t)) \tag{1}$$

Profit margin, μ_i, is updated using an adaptation rule based on the Widrow-Hoff "delta-rule with momentum" [23], as:

$$\mu_i(t+1) = \frac{(p_i(t) + \Gamma_i(t))}{\lambda_i} - 1 \tag{2}$$

$$\Gamma_i(t+1) = \gamma_i \Gamma_i(t) + (1 - \gamma_i)\Delta_i(t) \tag{3}$$

$$\Delta_i(t) = \beta_i(\tau_i(t) - p_i(t)) \tag{4}$$

where β_i is the *learning rate*, γ_i is *momentum coefficient*, and $\tau_i(t)$ is the current *target value*, calculated using the last

TABLE I: Glossary of acronyms.

Acronym	Description
SIP	Securities Information Processor
NBBO	National Best Bid and Offer (* indicates *next* NBBO value)
LOB	Limit Order Book (bids/offers ordered by price asc./desc.)
CDA	Continuous Double Auction
DCA	Discrete-time Call Auction
ZIC	Zero-Intelligence Constrained (background trader algorithm)
ZIP	Zero-Intelligence Plus (background trader algorithm)
LA	Latency Arbitrageur
2M	2 Market model (3M: 3 Market model, etc.)

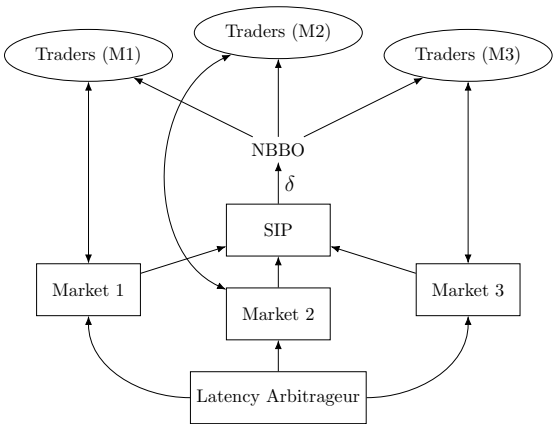


Fig. 2: Model containing three markets (3M). Each group of background traders have instant access to a primary market. SIP is updated with each new best bid and offer from all markets and, after delay, δ, the NBBO is published to all traders. The LA has instant access to all markets.

order price observed in the market, $q(t)$, plus or minus a small random deviation:

$$\tau_i(t) = R_i(t)q(t) + A_i(t) \tag{5}$$

$R_i(t)$ is a randomly generated coefficient that sets target price *relative* to $q(t)$; and $A_i(t)$ is a (small) random *absolute* price perturbation. When the intention is to increase price, $1.05 > R_i > 1$ and $A_i > 0$; to decrease price, $0.95 < R_i < 1$ and $A_i < 0$. For original ZIP implementation details, see [9, pp.41–45]. For detail on the effects of alternative ZIP implementations, refer to [24], [25]. In this study, learning rate β_i is independently sampled from a uniform distribution, such that $\beta_i \in U[0.1, 0.5]$. The momentum coefficient γ_i is sampled as: $\gamma_i \in U[0, 0.1]$. If the ith trader is a buyer, margin μ_i is initialised as $\mu_i \in U[-0.35, -0.05]$; for a seller, initial $\mu_i \in U[0.05, 0.35]$.

IV. Simulation Model Methodology

Here, we introduce the agent-based model used in this study.[1] Refer to Table I for a glossary of acronyms used throughout.

[1]For additional details on model, experiments, and results refer to [26]. Model code publicly available at https://github.com/mduffin95/MarketSim

A. Multiple Markets Model

The multiple markets model implemented in this study is an extension of the design introduced by Wah and Wellman [7]. Fig. 2 presents a schematic of an instantiation of the model containing three markets (or *exchange venues*). We refer to this as a 3M model (likewise, 4M has four markets, etc.).

The model contains a population of background traders, representing normal retail and institutional investors who do not employ high frequency trading (HFT) strategies. Traders are evenly distributed into three pools, with each pool assigned access to a primary market. Traders have direct and instantaneous access to their primary market, and can also view the National Best Bid and Offer (NBBO) across all markets. If NBBO prices are better than prices on the primary market, a trader can select to route an order to other markets to take advantage of the better price. The NBBO is calculated and updated by the Securities Information Processor (SIP). SIP monitors all markets and updates NBBO every time there is a change to a market's best bid and offer. The time taken for SIP to calculate and publish the new NBBO (denoted NBBO*) is represented by delay, δ. When $\delta = 0$, the NBBO always accurately reflects the latest prices at each market. Since traders can route their orders to any market, $\delta = 0$ is therefore equivalent to a single centralised marketplace. However, when $\delta > 0$, traders view a time-delayed NBBO that may contain *stale* prices that are no longer available at the underlying exchange. This effect is exacerbated as the value of δ is increased. Finally, the model allows for a latency arbitrageur (LA) to be positioned between the markets, with direct and instantaneous access to all. The LA represents HFT firms with low latency access to markets. Having instantaneous access, the LA is able to calculate its own private NBBO immediately. Therefore, when $\delta > 0$, the LA has a more up-to-date, and therefore more accurate view of the market than the background traders. The LA uses this advantage to perform arbitrage across the markets when the opportunities arise.

B. Latency Arbitrageur (LA)

The latency arbitrageur receives instantaneous pricing information from all of the markets that it is active in. Whenever it receives a new quote (BID, $OFFER$) from one of the markets it updates the recorded best bid/offer for that market. It then checks across all markets for an arbitrage opportunity. It does this by selecting the best bid (BID^*) and the best offer price ($OFFER^*$) across all markets and compares them. If $BID^* > (1+\alpha)OFFER^*$ then a large enough arbitrage opportunity exists. The variable α is used to create a minimum threshold, above which it is *worthwhile* to pursue the arbitrage opportunity (i.e., so that trading costs are covered). For all experiments, $\alpha = 0.001$.

C. Background Traders (ZIC / ZIP)

Background traders arrive to trade one unit at a rate determined by a Poisson process with rate λ (producing a mean interarrival time $1/\lambda$). Trader arrival is equivalent to an order arriving at an exchange, as it takes zero timesteps to send an order.

TABLE II: Default parameter values.

Parameter	Value	Description
α	0.001	Min profit threshold required for LA to trade
λ	0.075	Mean arrival rate of traders (Poisson process)
ρ	0.0006	Discounting factor for surplus calculation
$\bar{r}$	100 000	Mean fundamental value ($\bar{r} \equiv P_0$ in ZIP exps)
κ	0.05	Degree of fundamental price reversion to mean
R	2000	Range of admissible bid shading values (ZIC)
δ	0–1000	Delay in calculating the NBBO (latency)
T	250	Total number of background traders (ZIC/ZIP)
N	1000	Number of repeated trials per model setting (graphs plot mean $\pm 95\%$ confidence interval)

Following the model presented in [27], the surplus of background traders is discounted at a rate determined by ρ. Surplus before discounting is calculated as the difference between a trader's private valuation and the trade price: for buyers $PV_i - p_t$; and for sellers $p_t - PV_i$, where p_t is trade execution price, and PV_i is the private valuation of the ith trader. Discounting reduces the surplus of executed orders by a factor that increases the longer it takes for an order to execute. This represents the preference of traders for their orders to execute quickly.[2] For a transaction that executes at time t, between two orders that arrive at times $t(i)$ and $t(j)$, total discounted surplus is calculated as:

$$S = e^{-\rho(t-t(i))}(PV_i - p_t) + e^{-\rho(t-t(j))}(p_t - PV_j) \quad (6)$$

In the replication study (Section V), private values are assigned by perturbing a continuously and randomly varying fundamental value with mean $\bar{r} = 100\,000$. The parameter κ is used to control the degree to which the fundamental value reverts back to $\bar{r}$ (see Table II for full list of parameter values used). For the extension study (Section VI), we switch to using a symmetric supply and demand schedule to assign private values. Values are assigned between 63 000 and 137 000 in steps of 1000, thus giving a fixed equilibrium price $P_0 = 100\,000$. This approach simplifies performance metric calculations.

V. Model Validation: Replication Results

To validate our model implementation, we first present a strict replication of Wah and Wellman's results. Fig.3 plots total discounted surplus, S, against latency, δ, for a 2M model with and without LA present; and also for a single market (1M) with a central CDA. For 2M markets with LA present, surplus for background zero-intelligence traders only (ZIC only) and the combined market surplus of the background traders and LA (ZIC+LA) are plotted. When $\delta = 0$, the NBBO is always up-to-date so no arbitrage opportunities emerge and ZIC orders are always routed to the correct exchange. Therefore, surplus is equal for all models ($S \approx 2.95 \times 10^6$). Total market surplus in the central CDA ($S_{\text{CDA}} \approx 2.95 \times 10^6$) is invariant to δ as no arbitrage opportunities exist. In the 2M model with LA present, total surplus tends to a lower value ($\lim_{\delta\to\infty} S_{\text{ZIC+LA}} \approx 2.8 \times 10^6$). The surplus (profit) gained by the LA is the difference between the curves ZIC+LA and ZIC only. Finally, in the 2M model with no LA, surplus is highest ($\lim_{\delta\to\infty} S_{\text{ZIC}} \approx 3 \times 10^6$). Non-overlapping confidence intervals indicate that these differences are statistically significant. These results are quantitatively very similar to those presented by Wah and Wellman (compare Fig. 3 with [7, Fig. 5]) and provide strong validation of correct model implementation.

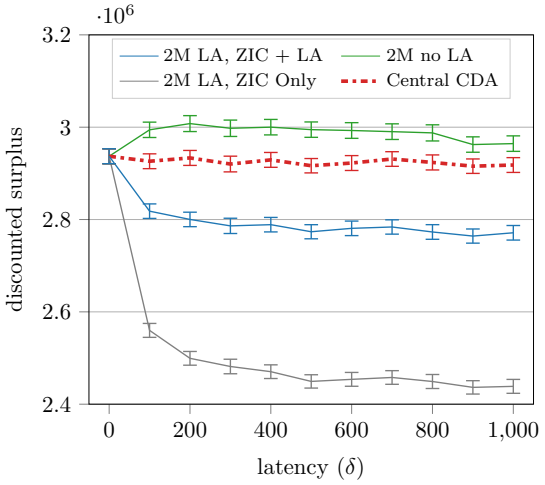


Fig. 3: Replication validation. Total discounted surplus, S, against δ in 2M model, with/without LA, and a central CDA.

A. Conclusion: fragmentation benefits markets

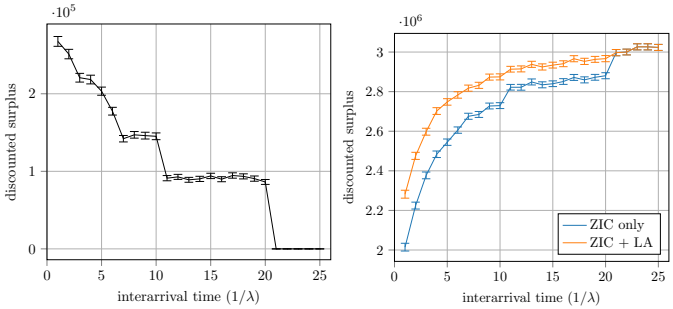
The results presented in Fig. 3 show that a fragmented 2M market with no LA (i.e., a market of ZIC traders only) produces more surplus than a centralised CDA; i.e., in a two market model, when $\delta > 0$, $S_{\text{ZIC}} > S_{\text{CDA}} > S_{\text{ZIC+LA}}$. Also observing this effect, Wah and Wellman [7] conclude:

> "...for continuous markets, fragmentation can actually provide a benefit, as the separated markets are less likely to admit inefficient trades (i.e., where both traders' values fall on the same side of the longer-term equilibrium price) that arise due to the vagaries of arrival sequences."

We further investigate this unexpected and counterintuitive result by increasing fragmentation in the model. Specifically, we observe total discounted surplus, S, as number of markets is increased (2M, 3M,..., 5M). Results (not shown)[3] demonstrate that greater fragmentation leads to greater LA profits. This is expected, given that greater fragmentation (more markets) is likely to lead to a greater number of arbitrage opportunities. However, we also see that in markets both with and without an LA, total surplus, S, continues to increase as fragmentation is increased. This is unexpected, and provides further evidence for the counterintuitive conclusion

[2]Technically, discounting is also applied to the LA's orders, but since these always execute instantaneously (because the LA only submits orders when they are guaranteed to execute) the discount is always equal to zero.

[3]Results are presented and discussed fully in [26, Section 4.2].



(a) S_{LA} ($\delta = 20$) (b) S_{ZIC} and $S_{\text{ZIC+LA}}$ ($\delta = 20$)

Fig. 4: Discounted surplus, S, against interarrival time of traders, $1/\lambda$. LA profits are eliminated when $1/\lambda > \delta = 20$.

that fragmentation benefits markets. We return to this, and investigate more fully, in Section VI. First, we look at the latency bounds on arbitrage opportunities in fragmented markets.

B. Arrival rate and latency: an upper bound on arbitrage

Here, we explore the relationship between arrival rate of traders (λ; more specifically, interarrival time $1/\lambda$) and NBBO latency, δ, using a 2M model. Arrival rates are set to a constant value λ (rather than generated by a Poisson process with mean λ, as previously), and latency is fixed to $\delta = 20$. Fig. 4a shows a stepped decline in latency arbitrageur surplus, S_{LA}, as the interarrival time of traders, $1/\lambda$, is increased.[4] We see that $S_{\text{LA}} = 0$ when $1/\lambda > \delta = 20$. Therefore, if NBBO delay δ can be restricted to values below mean interarrival time of traders, $1/\lambda$, then LA profits can be eliminated—i.e., latency arbitrage is not possible if NBBO delays can be kept below strict upper bound $1/\lambda$. In the real world, however, this is unlikely to be achievable in practice; particularly in liquid markets, where $1/\lambda \ll 1$ms and falling rapidly.[5]

In Fig. 4b, we see that total discounted surplus of traders increases as interarrival time increases. At low interarrival times (i.e., fast arrival rates), surplus is lowered due to poor order routing based on SIP's publication of stale NBBO information. We return to order routing in Section VI-B.

VI. Results: Model Investigation

In this section we extend the model. First, we investigate Wah and Wellman's conclusion that fragmentation *benefits* markets (see Section V-A), by exploring the effects of bid shading (Section VI-A). We then introduce ZIP traders into the model (Section VI-C) and show that the results for central CDA, central call (DCA), and fragmented markets are very different to the results for models containing ZIC traders.

Henceforth, to simplify analysis, we alter the method of allocating private values to traders. Rather than allocating private values around a continually varying fundamental value with mean $\bar{r} = 100\,000$, traders are allocated private values according to a fixed supply and demand schedule with equilibrium price $P_0 = 100\,000$ (see Section IV-C for details). This alteration reduces total discounted surplus (for a central CDA model containing background ZIC traders $S_{\text{CDA}} \approx 2.05 \times 10^6$, rather than $S_{\text{CDA}} \approx 2.95 \times 10^6$ reported in Section V). However, all observations reported in Section V remain consistent—the methodological change introduced to private value allocation of traders does not otherwise affect results, apart from reducing S under all settings.

A. Bid shading and efficiency

Fig. 5 shows results of varying R in a 2M model containing ZIC traders. Three latency settings are plotted: $\delta = 0$ (equivalent to central CDA), $\delta = 100$, and $\delta = 1000$. In Fig. 5a, total discounted surplus, S, is plotted. For most of the range of R values shown, we see that surplus for the CDA is highest (i.e., $S_{\delta=0} > S_{\delta=100} > S_{\delta=1000}$). Only at very low values of R (right, zoomed region) does this ordering change. Specifically, when $R = 2000$—the value used by Wellman and Wah [7] and replicated in Section V—the CDA produces the lowest surplus. This demonstrates that Wah and Wellman's conclusion that fragmentation *benefits* markets (Section V-A) is driven by a simulation artefact resulting from the selection of bid shading parameter, R. For values of $R \geq 12\,000$, we reach the converse conclusion: fragmentation *negatively* affects markets.

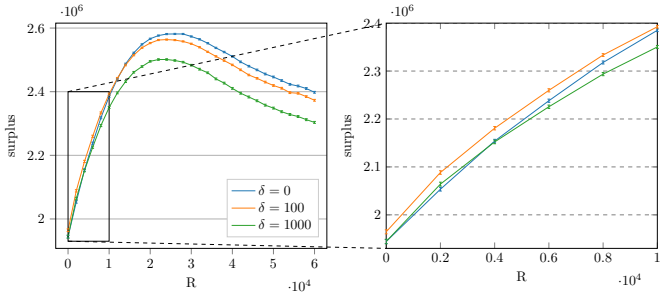
As R varies from low to high (Fig. 5a), S increases rapidly at first, reaches a maximum slightly below $R = 30\,000$, and then declines more gradually. We can better understand this by considering orders that are traded in the market. Fig. 5b plots the fraction of trades that involve an *extramarginal* (EM) trader.[6] We see that at low values of R, the proportion of trades involving an EM trader is well above 0.6. This fraction falls as R increases, tending to a value of 0.1 when $R \geq 40\,000$. Therefore, if we consider R in relation to competitive equilibrium, $P_0 = 100\,000$, we see that maximum surplus is generated when $R/P_0 \approx 0.3$, and inefficient trades fall to lowest levels when $R/P_0 \approx 0.4$. These results align well with those previously shown by Zhan and Friedman [28, Fig. 6]. In particular, Zhan and Friedman showed that CDA markets generate greatest surplus when traders markup (i.e., bid shade) by 30%; while surplus loss due to EM traders executing orders tends to zero when markup is greater than 40%.[7] Above $R/P_0 \approx 0.3$, total surplus falls because excessive bid shading stops some *intramarginal* (IM) traders from making trades

[4] S_{LA} decreases in discrete steps due to the way that interarrival time, $1/\lambda$, divides into NBBO latency, $\delta = 20$. Interarrival times between 7 and 10, inclusive, divides into 20 twice (e.g. $2 \equiv 7 \mod 20$), so each produce the same discounted surplus; similarly for $1/\lambda \in [11..20]$.

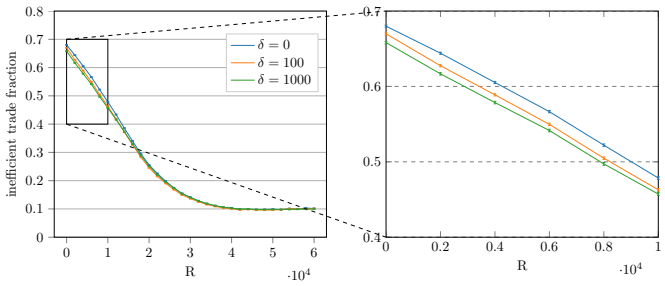
[5] See Nanex's animation of a 10ms period of trading in symbol MRK on May 16, 2013: https://www.youtube.com/watch?v=L5cZaIZ5bWc; and associated report http://www.nanex.net/Research/IsNBBOIgnored.html.

[6] EM traders are buyers with private value below competitive equilibrium, $v_b < P_0$, and sellers with private value $v_s > P_0$. If markets trade consistently at competitive equilibrium price, P_0, then EM traders are unable to trade.

[7] In Zhan and Friedman's model [28], markup for trader i is calculated as a fixed percentage of trader i's private value, v_i. In our model, ZIC traders randomly select the amount to bid shade (markup) from a uniform distribution around v_i bounded by R. Therefore, in our model inefficient trades tend to a background level of 10%, rather than fall to zero, as EM trades are still possible even when values of R are large.



(a) Total surplus, S, for three different values of δ. Central CDA is equivalent to $\delta = 0$. For the replication results (Fig. 3), $R = 2000$; at this value (right, zoomed region), $S_{\delta=100} > S_{\delta=1000} > S_{\delta=0}$.



(b) Inefficient trade fraction (inefficiently traded orders / total fulfilled orders). Notice that $\delta = 0$ (i.e., central CDA) has the highest fraction (≈ 0.65) of inefficient trades when $R = 2000$ (right, zoomed region).

Fig. 5: (a) Total surplus, S, and (b) inefficient trade fraction for ZIC markets, plotted against bid shading value, R.

they otherwise could make.[8] In the search for profit, an IM trader that shades private value by an excessively large amount will be unable to trade at competitive equilibrium price, P_0. This reduces total market surplus.

B. Order routing and latency

In Fig. 5b, when $R = 2000$, we see that the fraction of EM trading reduces as latency δ is increased. To understand this, we consider how latency affects order routing in the 2M model. As latency is increased from $\delta = 0$ to $\delta = 1000$, the proportion of orders routed incorrectly (i.e., routed to an exchange based on a *stale* price in the NBBO that is no longer available) grows quickly from zero and settles around 20%.[9] When an order is incorrectly routed, the price available upon arrival is *a priori* unknown, and determined by the arrival of new orders during the latency period. As intramarginal (IM) traders are more likely to submit a new best price, then a poorly routed order effectively offers the opportunity of an "extra roll of the die" to transact efficiently with an IM trader. Although relatively small, this effect (when $R = 2000$) is enough to lead to the conclusion that fragmentation benefits markets, presented in the replication results (see Section V).

C. Introducing ZIP traders

We have clearly demonstrated that the bid shading value, R, has a profound affect on model results containing ZIC traders. In particular, the value $R = 2000$ used in the replication study (Section V) leads directly to the counterintuitive conclusion that fragmentation benefits markets. To overcome model reliance on parameter R, we now introduce ZIP traders. Each ZIP trader, i, individually and autonomously adapts bid-shading margin, μ_i, during trading. This is a more realistic model of human trading and removes the need for R.

Fig 6 shows total surplus for 2M markets containing ZIP traders. When $\delta = 0$, equivalent to central CDA, we see $S^{\text{ZIP}}_{\text{CDA}} \approx 2.57 \times 10^6$. Under the new private value allocation, this is higher than the total surplus for a central CDA in ZIC

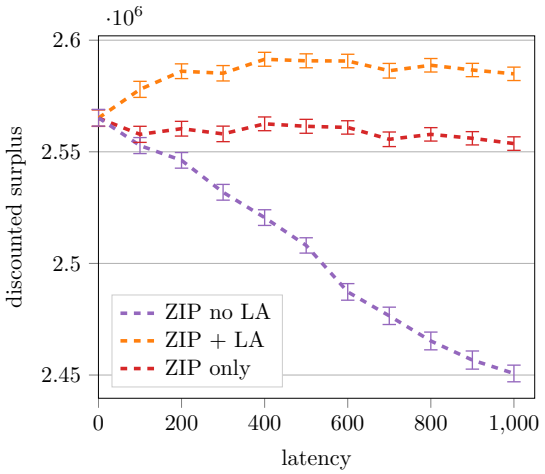


Fig. 6: Surplus for ZIP markets. $S^{\text{ZIP}}_{\text{CDA}} \approx 2.57 \times 10^6$.

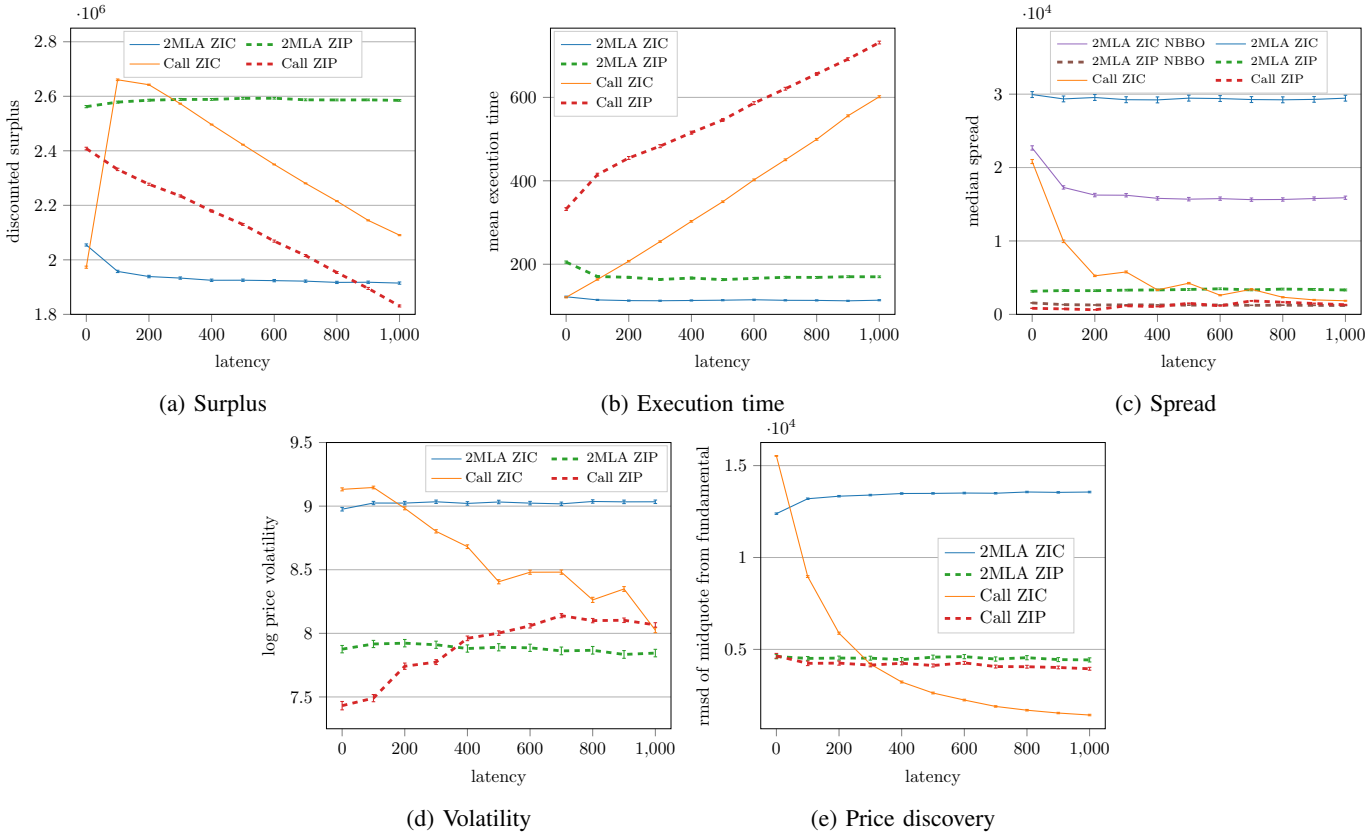
markets $S^{\text{ZIC}}_{\text{CDA}} \approx 2.05 \times 10^6$. In a fragmented market with no LA, surplus (S_{ZIPnoLA}) falls linearly as δ increases. This is in complete contrast to results in ZIC markets (compare Fig. 3), where surplus in fragmented markets is higher than a central CDA. In 2M markets with LA present, we see that surplus for ZIP traders only (S_{ZIPonly}) decreases very slightly as latency increases; and total surplus for ZIP traders and LA ($S_{\text{ZIP+LA}}$) is greater than for a central CDA. This suggests that latency arbitrageurs *benefit* background traders in a fragmented market, by counteracting the effects of latency, while generating additional surplus not otherwise available in a central CDA. Again, this result contrasts with ZIC markets (Fig. 3), where LA presence is shown to negatively affect background traders and the market overall.

D. Discrete-time call auction (DCA)

Here, we introduce call auction (DCA) markets and compare results with CDA markets (Fig. 7) for models containing ZIC (solid lines) and ZIP (dashed lines) traders. For DCA markets, latency on the x-axis refers to the *clearing interval* of the market; for CDA markets latency refers to NBBO delay, δ.

[8] IM traders are buyers with private value above competitive equilibrium, $v_b \geq P_0$, and sellers with private value $v_s \leq P_0$. If markets trade consistently at competitive equilibrium price, P_0, then IM traders are able to trade.

[9] For detail on order routing and efficiency, see [26, Section 4.42, Fig. 4.7].



(a) Surplus (b) Execution time (c) Spread

(d) Volatility (e) Price discovery

Fig. 7: Comparison between CDA and call auction (DCA), for models containing ZIC (solid line) and ZIP (dashed) traders.

Results for ZIC markets accurately replicate those presented by Wah and Wellman [7, Fig. 7]. The only difference: here, RMSD falls exponentially in ZIC call markets (Fig. 7e) due to using a fixed equilibrium value P_0; when using a continuously varying fundamental value with mean $\bar{r}$, RMSD remains high for all clearing intervals. When using fixed equilibrium, the call market is able to aggregate information from all collected orders to discover the equilibrium. This is not possible when the equilibrium price continuously changes.

Let us consider call market results in turn:

Surplus (Fig. 7a) ZIC: call markets increase surplus as long as clearing time is not too long; ZIP: call markets reduce surplus—they do not help.

Execution time (Fig. 7b) As expected, call markets increase execution time: increasing linearly with clearing time. Execution times are higher for ZIP than ZIC (this is expected, as ZIP waits to trade for higher profit).

Spread (Fig. 7c) ZIC: call markets reduce spread. ZIP: spreads are much lower for all markets (as traders *intelligently* post orders close to equilibrium), with call market spreads similar to 2M NBBO spreads with LA present.

Volatility (Fig. 7d) Volatility is lower in ZIP markets. As clearing time increases, volatility falls in ZIC markets and increases in ZIP markets. Shorter DCA clearing times are best for ZIP; longer clearing times best ZIC.

Price discovery (Fig. 7e) Price discovery is much better in ZIP markets (i.e., lower RMSD). Little difference in price discovery between CDA and call market for ZIP traders.

Overall, for ZIP traders, CDA is better than DCA for: surplus (7a) and execution time (7b); similar for spread (7c) and price discovery (7e); and worse for volatility at low latencies, better at high latencies (7d). Therefore, CDA outperforms DCA when traders are modelled using ZIP.

For ZIC traders, DCA is better than CDA for: surplus (7a), spread (7c), and volatility at longer clearing intervals (7d); and worse for execution time (7b) and price discovery (7e). Therefore, there is some evidence for DCA outperforming CDA when traders are modelled using ZIC (a similar finding to that presented by Wah and Wellman [7, Fig. 7]).

E. Discussion: The Rise of Periodic Auctions

On 3 Jan 2018, MiFID II regulation came into force, requiring European trading venues to have greater trading transparency. As a result, there has been a migration in trading volume from entirely opaque dark pool venues, to a host of new semi-transparent periodic auction venues (equivalent to DCA in the model). Current debate centres around whether periodic venues have enough transparency of information to enable algorithms to trade efficiently, whether there is enough price discovery, and whether they enable unfair opportunities for brokers to self-match. Regulators, such as the FCA in UK, are currently investigating [29]. This demonstrates the timeliness

of research into discrete-time call auctions, and the results we present suggest that the lack of transparency during the call phase does indeed reduce efficiency in markets (Section VI-D). Like real-world automated trading systems, ZIP algorithms require up-to-date price information in order to efficiently trade. The lack of order transparency during the auction phase reduces the ability of ZIP traders to optimise profit margins.

VII. Conclusions

We have presented a strict replication, investigation, and extension of Wah and Wellman's [7] agent-based model of latency arbitrage in a fragmented market. Using a population of zero-intelligence (ZIC) trading agents, Wah and Wellman's results led them to conclude that "market efficiency is negatively affected by the actions of a latency arbitrageur", while "a discrete-time call market... eliminates latency arbitrage opportunities and improves efficiency." However, we have demonstrated that results from models using ZIC traders are highly sensitive to the value selected for bid shading (alternatively called *markup*, or *margin*) parameter R; and the value used by Wah and Wellman ($R = 2000$; only 2% of the value of market equilibrium price) is inefficient for continuous double auction (CDA) markets, enabling a high proportion of extramarginal trading (see Fig. 5). This raises doubt on the validity of Wah and Wellman's conclusions. We have also shown an upper bound on the ability to perform latency arbitrage in fragmented markets (see Fig. 4). If NBBO delays can be restricted below the mean interarrival time of orders in the market, then latency arbitrage can be eliminated. In practice, however, this is extremely difficult—perhaps impossible—to implement in liquid real-world markets, where interarrival times can be much less than 1ms.

Most importantly, we have shown that switching to a more realistic model of trading behaviour (zero-intelligence plus; ZIP—traders that *intelligently* adapt profit margin based on the prevailing market conditions), results are such that: (i) fragmented markets *benefit* from LA intervention (see Fig. 6); and (ii) discrete-time call markets (DCA) *do not* improve efficiency when compared with CDA markets (see Fig. 7).

These contrary conclusions indicate the importance of introducing computational intelligence into trading strategies when modelling financial markets. We propose that a population of ZIP agents more accurately models the behaviour of financial markets than a population of ZIC agents, and therefore our results are more likely to be indicative of the real world. However, there remain plenty of avenues of exploration to perform: (i) understanding how variations on the discrete-time call auction affects efficiency; (ii) studying the effects of different order types, such as inter-market sweep orders (IMSO); (iii) studying HFT strategies that rely on short-term price predictions rather than latency arbitrage. We reserve these investigations for future work.

References

[1] M. Lewis, *Flash Boys: A Wall Street Revolt*. Norton & Co., 2014.

[2] D. MacKenzie, "A sociology of algorithms: High-frequency trading and the shaping of markets," June 2014, Univ. of Edinburgh, Working Paper.

[3] M. Durbin, *All about high-frequency trading*. McGraw-Hill, NY, 2010.

[4] D. Schneider, "The microsecond market," *IEEE spectrum*, vol. 49, no. 6, pp. 66–81, 2012.

[5] E. Wah, "How prevalent and profitable are latency arbitrage opportunities on U.S. stock exchanges?" Feb 8 2016, available at SSRN: http://ssrn.com/abstract=2729109.

[6] R. P. Bartlett and J. McCrary, "How rigged are stock markets? Evidence from microsecond timestamps," 1 May 2017, UC Berkeley Public Law Research Paper No. 2812123. Available at SSRN: https://ssrn.com/abstract=2812123.

[7] E. Wah and M. P. Wellman, "Latency arbitrage, market fragmentation, and efficiency: a two-market model," in *Proc. 14th ACM Conf. on Electronic Commerce*, Philadelphia, PA, June 2013, pp. 855–872.

[8] D. K. Gode and S. Sunder, "Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality," *J. Political Economy*, vol. 101, no. 1, pp. 119–137, 1993.

[9] D. Cliff, "Minimal-intelligence agents for bargaining behaviors in market-based environments," Hewlett-Packard Labs., Tech. Rep. HPL-97-91, Aug. 1997. [Online]. Available: http://bit.ly/18uC9vM

[10] R. Das, J. Hanson, J. Kephart, and G. Tesauro, "Agent-human interactions in the continuous double auction," in *17th Int. Joint Conf. Artif. Intell. (IJCAI-01)*, Seattle, USA, Aug. 2001, pp. 1169–1176.

[11] J. Hasbrouck and G. Saar, "Low-latency trading," *J. Financial Markets*, vol. 16, no. 4, pp. 646–679, 2013.

[12] A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, "The flash crash: High-frequency trading in an electronic market," *J. of Finance*, vol. 72, no. 3, pp. 967–998, 2017.

[13] M. Baron, J. Brogaard, and A. Kirilenko, "The trading profits of high frequency traders," *Unpublished Manuscript*, 2012.

[14] S. N. Cohen and L. Szpruch, "A limit order book model for latency arbitrage," *Mathematics and Financial Economics*, vol. 6, no. 3, pp. 211–227, 2012.

[15] T. A. Hanson, "The effects of high frequency traders in a simulated market," in *Midwest Finance Associated Annual Meetings Paper*, 2012.

[16] R. A. Jarrow and P. Protter, "A dysfunctional role of high frequency trading in electronic markets," *Int. J. Theoretical and Applied Finance*, vol. 15, no. 03, article 1250022, 2012.

[17] H. Mendelson, "Consolidation, fragmentation, and market performance," *J. Financial and Quantitative Analysis*, vol. 22, no. 2, pp. 189–207, 1987.

[18] S. Ding, J. Hanna, and T. Hendershott, "How slow is the NBBO? a comparison with direct exchange feeds," *Financial Review*, vol. 49, no. 2, pp. 313–332, 2014.

[19] E. Wah and M. P. Wellman, "Latency arbitrage in fragmented markets: A strategic agent-based analysis," *Algorithmic Finance*, vol. 5, no. 3-4, pp. 69–93, 2016.

[20] E. Wah, D. Hurd, and M. P. Wellman, "Strategic market choice: Frequent call markets vs. continuous double auctions for fast and slow traders," *EAI Endorsed Trans. Serious Games*, vol. 3, p. e1, 2016.

[21] J. Cartlidge, C. Szostek, M. D. Luca, and D. Cliff, "Too fast too furious: faster financial-market trading agents can give less efficient markets," in *Proc. 4th Int. Conf. on Agents and Artificial Intelligence (ICAART), Vol. 2*, Vilamoura, Portugal, Feb 2012, pp. 126–135.

[22] J. Cartlidge and D. Cliff, "Exploring the robot phase transition in experimental human-algorithmic markets," Foresight Report–The Future of Computer Trading in Financial Markets, Driver Review, DR25, London: UK Government Office for Science, Tech. Rep., Apr 2012.

[23] B. Widrow and M. E. Hoff, Jr., "Adaptive switching circuits," *Inst. Radio Engineers, Western Electron. Show and Conv. (IRE WESCON), Conv. Rec.*, vol. 4, pp. 96–104, Aug. 1960.

[24] S. Stotter, J. Cartlidge, and D. Cliff, "Exploring assignment-adaptive (ASAD) trading agents in financial market experiments." in *Proc. 5th Int. Conf. on Agents and Artificial Intelligence (ICAART), Vol. 1*, Barcelona, Spain, Feb 2013, pp. 77–88.

[25] ——, "Behavioural investigations of financial trading agents using exchange portal (ExPo)," in *Transactions on Computational Collective Intelligence XVII*. Springer, Berlin, Nov 2014, pp. 22–45.

[26] M. J. Duffin, "Investigating the effects of latency arbitrage on fragmented markets," Master's thesis, Dept. Comp. Sci., Univ. of Bristol, June 2018.

[27] R. L. Goettler, C. A. Parlour, and U. Rajan, "Informed traders and limit order markets," *J. Financial Economics*, vol. 93, no. 1, pp. 67–87, 2009.

[28] W. Zhan and D. Friedman, "Markups in double auction markets," *J. Economic Dynamics and Control*, vol. 31, pp. 2984–3005, 2007.

[29] Financial Conduct Authority, "Periodic auctions," Available: https://www.fca.org.uk/publications/research/periodic-auctions, Jun 2018.