BANK OF ENGLAND

Staff Working Paper No. 743

The deeds of speed: an agent-based model of market liquidity and flash episodes

Geir-Are Kårvik, Joseph Noss, Jack Worlidge and Daniel Beale

July 2018

BANK OF ENGLAND

Staff Working Paper No. 743
The deeds of speed: an agent-based model of market liquidity and flash episodes

Geir-Are Kårvik,[1] Joseph Noss,[2] Jack Worlidge[3] and Daniel Beale[4]

Abstract

This paper examines the role of high-frequency traders in flash episodes in electronic financial markets. To do so, we construct an agent-based model of a market for a financial asset in which trading occurs through a central limit order book. The model consists of heterogeneous agents with different trading strategies and frequencies, and is calibrated to high-frequency time series data on the sterling-US dollar exchange rate. Flash episodes occur in the model due to the procyclical behaviour of high-frequency market participants. This is aligned with some empirical evidence as to the drivers of real-world flash crashes. We find that the prevalence of flash episodes increases with the frequency with which high-frequency market participants trade compared to their low-frequency counterparts. This provides tentative theoretical evidence that the recent growth in high-frequency trading across some markets has led to flash episodes. Furthermore, we adapt the model so that large movements in price trigger temporary halts in trading (ie circuit breakers). This is found to reduce the magnitude and frequency of flash episodes.

Key words: Agent-based modelling, high-frequency trading, financial stability, market liquidity, flash episodes, principal trading firms (PTFs).

JEL classification: C63, G11, G12, G17.

(1) Bank of England. Email: geir-are.karvik@bankofengland.co.uk
(2) Bank of England. Email: joseph.noss@bankofengland.co.uk
(3) Bank of England. Email: jack.worlidge@bankofengland.co.uk
(4) Bank of England. Email: daniel.beale@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to Matthew Allan, Evangelos Benos, Robert Hillman, Zijun Liu, Grainne McGread, Louise Otter, Lucas Pedace, William Rawstorne, Jon Relleen, Rhiannon Sowerbutts, Arthur Turrell, Edward White and participants at an internal Bank of England seminar for helpful discussion and comments. We are grateful to Thomson Reuters for providing data.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Publications and Design Team, Bank of England, Threadneedle Street, London, EC2R 8AH
Telephone +44 (0)20 7601 4030 email publications@bankofengland.co.uk

ISSN 1749-9135 (on-line)

1. Introduction

There have been a number of recent 'flash episodes' in major financial markets. Such episodes consist of large and rapid changes in the traded price of an asset that do not coincide with – or substantially overshoot – changes in economic fundamentals. Several such events have occurred in markets that are among the largest and most liquid in the world. These include US equities (6 May 2010), US Treasuries (15 October 2014), and the sterling-US dollar exchange rate (7 October 2016).[1]

Flash episodes have caught the interest of policymakers. This is unsurprising given the role that liquid smooth-functioning markets play in a well-functioning financial system (see Bank of England (2017)). Financial Markets in which price changes are orderly and reflect changes in valuation factors are desirable in that they allow for the orderly matching of buyers and sellers.

Flash episodes in major financial markets have so far been short-lived and have not, as yet, had immediate consequences for financial stability. But flash episodes could potentially pose such risks if they were to undermine investor confidence in the price at which securities could be transacted, and in doing so impede investment. This could happen if flash episodes were to become more frequent or if market disruption were to become longer-lasting. It is therefore important to understand how such episodes arise, and potential policy tools that might mitigate them.

Markets affected by recent major flash episodes have a high degree of algorithmic trading, where trading and execution decisions are automated and made electronically by computer programs (algorithms). This includes algorithmic trading operating at high frequency. And the presence and activity of high-speed algorithmic trading (or, high-frequency trading, as we henceforth refer to it) has grown over the last 15 years to almost three quarters of trading volume in some markets (Moore, Schrimpf and Sushki (2016)).

Empirical evidence suggests that during periods of extreme stress high-frequency traders might exacerbate changes in prices by behaving procyclically (see SEC/CFTC (2010), IMF (2015), Bank of England (2016), BIS (2017) and Kirilenko et. al (2017)). Such procyclical behaviour can include (i) temporarily exiting markets when functioning starts to break down, reducing overall participation and liquidity; and/or (ii) exacerbating changes in price by selling (buying) into falling (rising) markets, undermining their liquidity. That said, none of these behaviours is unique to high-frequency traders; rather, they might represent the prudent actions of market participants – whatever their frequency of trading – under stressed conditions. But given the relatively recent increase in the importance of high frequency trading, there is still only limited understanding of the circumstances under which the behaviour of high frequency traders can combine under stress and give rise to flash episodes.

This paper aims to improve understanding of the role high-frequency traders play contributing to flash episodes. To do so, it forms an 'agent-based model' that draws together the behaviour of a number of different types of financial market participants and investigates how they can combine to lead to flash episodes. Agent-based models are a promising tool for understanding the dynamics of algorithmic trading, given how they consist of a number of distinct agents that follow-predetermined rules in a manner analogous to how algorithmic trading behaves in reality (see Turrell (2016)). This was recognised by Kim and Markowitz (1989) in their explanation of the 1987 'Black Monday' market crash.

[1] Other similar episodes include that on 15 January 2015 , the day of the removal of the peg of the Swiss Franc to the Euro; and 24 August 2015, a day of significant market dysfunction and volatility in US equity markets. See Massad (2015) for statistics on flash events in selected futures contracts and BIS (2017) for a list of flash episodes in FX markets.

This paper provides a microstructure model of a single security traded on a central limit order book in which market participants follow fixed behavioural rules. In doing so, it draws on similar agent-based models that have been used to study the dynamics of financial market liquidity, including those of Bookstaber and Paddrik (2015) and Braun-Munzinger et al (2016). As in Brewer et al (2013), flash crashes in this model are precipitated by the arrival of a large order. Such occasional large orders mean the model gives rise both to stable market prices under normal conditions, as well as sharp swings in price under stress - similar to the models of Leal et al (2014, 2017).

The unique contribution of this work, however, is the central role played in the model by market makers. Such market makers do not take a fundamental view on the value of assets, but instead hold inventories of securities with the aim of profiting from short-term imbalances in their supply/demand from other market participants through setting a bid-ask spread. Doing so exposes them both to the risk of incurring losses on their inventory due to changes in prices, and of the risk of being adversely selected by better informed market participants (Glosten and Milgrom (1985)). In stress, a perceived increase in both these risks causes market makers to withdraw their provision of liquidity. The dynamics of market liquidity – including the emergence of flash episodes – therefore arises endogenously via the interaction of agents that consume and provide liquidity. This matches the empirically observed withdrawal of liquidity observed in several flash events (see Section 2).

Our baseline model consists of five distinct agents who trade a single security via a central limit order book. Two of these market participants are market makers, who make quotes to buy and sell at a certain price (i.e. they supply liquidity), but do not themselves initiate transactions. Three other participants vary in their motivation for trading, but all initiate transactions with other agents (that is, they demand liquidity). This baseline model is calibrated to match the dynamics of high-frequency time series data on the sterling-dollar exchange rate, taken from the Thomson Reuters Matching platform. We find that our model is able to produce simulations of market prices whose high frequency dynamics do not differ significantly from those in the empirical data.

We also consider a second version of our model in which participants differ greatly in the frequency with which they observe, as well as trade in, the market. Those observing the market at a high frequency are analogous to real-world high-frequency trading strategies, and trade based on relatively simple heuristics. But one of the slower market participants trades on the basis of its belief as to the value of the asset that is commensurate with economic fundamentals.

Our findings are broadly three-fold:

First, when all agents observe – and have the opportunity to trade in – the market with similar frequencies, the above behaviours combine to yield times series of prices that are stable and continuous, despite the presence of agents with behavioural rules that are inherently procyclical.

But, when some of the agents trade at a markedly higher frequency than others, this gives rise to flash episodes. Such episodes arise due to the procyclical behaviours of high-frequency market participants, which combine to produce dynamics that roughly match those observed in real-world flash episodes. An initial movement in market price leads to a reduction in risk taking by high-frequency market makers, who withdraw their provision of liquidity in order to reduce the risk of being adversely selected by market participants with information about the true price of the asset. At the same time, other high-frequency traders consume liquidity by selling securities into already falling markets. This leads to a rapid and self-fulfilling reduction in price, which takes place at a horizon shorter than that over which lower-frequency traders can step in to return prices to a level commensurate with fundamentals.

Second, we find that the prevalence of flash episodes increases as we increase the difference in the frequency with which slower and faster-moving market participants trade. The greater the frequency of trading of high – compared to low – frequency market participants, the greater the potential for the above procyclical dynamics to take hold before slower moving market participants can arrest the resulting fall in price. This finding provides some tentative evidence to support the conclusion that it is the growth in high-frequency trading that has led to the recent increase in flash episodes.

Importantly, and in contrast to the findings of Brewer et al (2013), a large initial movement in price is not enough – in itself – to precipitate a flash episode. Rather, in our model, flash episodes arise only when agents with procyclical strategies trade at a higher frequency, which allows the dynamics they generate to take hold over a horizon shorter than that at which slower moving market participants can intervene to restore orderly price formation.

Finally, we use the model to investigate potential policies that might reduce the frequency and severity of flash episodes. Similar to Brewer et al (2013), we find that, according to the model set up and agent dynamics described here, the introduction of mandatory halts in trading (circuit breakers) has the effect of reducing the frequency of flash episodes. Such circuit breakers provide a pause during which lower-frequency market participants can observe any reduction in market prices below their perceptions of fundamental value, so that when trading resumes, they are able to arrest the fall in price by submitting countervailing buy orders.

We proceed as follows. Section 2 gives an overview of recent structural changes in selected financial markets, the emergence of flash episodes, and empirical evidence as to the behaviour of high-frequency traders therein. The model and its calibration is given in Section 3 and 4 respectively. Section 5 examines a simulation of prices from the baseline model, and shows how – when all participants are able to trade with similar frequency – they combine to result in stable price paths. Section 6 shows how the model can be used to simulate flash episodes. This section also gives an example of how the model can be used to examine the efficacy of potential policy interventions, including the application of circuit breakers. A final section concludes.

2. Structural changes in financial markets and the emergence of flash episodes

The proportion of electronic trading in financial markets has increased substantially over recent decades. This is particularly the case in markets with relatively simple and/or standardised securities, such as equities, foreign exchange, futures and some government bond markets (see Bank of England (2017)). This shift to electronic trading has allowed for greater transparency around the prices at which (at least some) market participants are able to transact (Salmon (2017)). This – combined with advances in technology and, in some markets, changes in regulation[2] – has led to an increase in algorithmic trading, where trading decisions and execution are fully automated, taking advantage of increased data availability.

The increase in algorithmic trading includes so-called high-frequency trading – which uses advanced technology to enhance information gathering and decision making, as well as the execution and routing of orders. High-frequency trading is used by a number of different types of market participants. But recent years have also seen the emergence of new specialist 'principal trading firms' (or PTFs) that trade at high frequency on their own account, often with a thinner degree of capitalisation and shorter holding periods than their more traditional low-frequency counterparts (Foucault (2016)).

[2] In equity markets, regulation has played a role in affecting the growth of other trading systems since the late 1990s (Anderson et. al (2015) and Salmon (2017)).

The growth of electronic and automated trading has also been accompanied by the emergence of flash episodes in some markets, including those that are traditionally very liquid. In such episodes, prices move sharply, before largely reversing – all within a matter of minutes, and to a degree that far exceeds changes in perceptions of economic fundamentals. Markets that have seen flash episodes are predominantly traded via central limit order books. Such limit order books provide a mechanism through which a number of participants can submit orders to trade at a variety of prices. Such orders comprise, broadly, of two types:

- 'Limit orders', which consist of an expression of commitment to trade at a specific price in a given volume and direction. In doing so, they typically *supply* liquidity, as they allow others to initiate transactions against them if they so wish.

- 'Market orders', which initiates a trade against an existing limit order(s). Market orders to buy (sell) are typically against limit orders to sell (buy) that are lowest (highest) in price across the order book. Market orders are typically seen to *demand* liquidity, as they immediately initiate transactions at prices at which they have offered to trade.[3]

The sharp movements in price seen during flash episodes consist of large imbalances between the supply and demand for liquidity – that is, a quantity of market orders to sell that depletes the available limit orders to buy.

One reason for the observed reduction in the supply of liquidity during flash episodes might be market makers' fear of adverse selection – that is, their perceived risk of transacting with market participants that are party to superior, or more up-to-date, information.[4] That could arise due to a market order (or orders) that are large relative to the supply of available limit orders. This could be interpreted by market makers as a signal that their offered prices do not reflect other participants' more up-to-date assessment of economic fundamentals, and that prices could therefore move against them. If so this might cause them to withdraw their supply of liquidity. This dynamic was seen at the start of the 2014 flash episode in US treasuries, where a large market order led to marked reductions in liquidity supply (IMF (2015)).

Perceived risk of adverse selection can also arise as a result of an unanticipated shock to – or increase in uncertainty around – economic fundamentals. This can also lead market makers to withdraw liquidity to avoid the risk of trading with better informed participants. Such a dynamic arose immediately following the depegging of the Swiss Franc to Euro. Immediately following the announcement of the depeg, liquidity supply contracted significantly, leading to a sharp fall in the exchange rate, part of which subsequently reversed (Cielinska et. al (2017)).

Existing literature also suggests that flash episodes might be exacerbated by an increase in demand for liquidity. This might come about as a result of:

(i) Directional trading strategies that attempt to profit from short-term price trends and demands liquidity in the same direction as recent price moves. For example, high-frequency traders' net aggressive selling of E-mini futures increased markedly during the US equity flash crash in 2010 (SEC/CFTC (2010)).

(ii) Dealers' hedging of their exposure to options positions, which can lead them to buy/sell large quantities of securities as their prices rise/fall, in order to maintain a neutral position with respect to further price movements. Such behaviour was seen during the 2016 flash episode in the Sterling-US dollar exchange rate (see BIS (2017)).

[3] More precisely, market orders will be executed and result in a trade if there is a sufficient quantity of limit orders at the same price. Limit orders will be executed and result in a trade either as a result of their being matched with a market order, or a limit order, at the same price level but with the opposite trading interest.
[4] See Glosten and Milgrom (1985), Kyle (1985) and Easley and O'Hara (1987).

(iii) Client orders to which dealers are committed to executing mechanistically when prices fall beyond a certain pre-determined level (eg. 'stop-loss' orders). This form of procyclical selling was also seen during the 2016 sterling flash episode (see BIS (2017)).

What remains unclear is how and why these behaviours occasionally lead to flash episodes. After all, none of the incentives to withdraw or consume liquidity described above are specific to algorithmic market making. Rather, they could represent rational and prudent behaviour on an individual level by any market participant seeking to manage its risk or maximise its profitability, whether it be algorithmic or human.

In the next section we seek to answer this question. We do so by forming an agent-based model of a market for a single security, where agents' behaviours roughly mirror those described above.

3. The model

The five agents in the model are divided into four types. These differ in their motives for trading:

(i) Two competing market makers supply liquidity by submitting limit orders, against which other participants can trade.[5] In the version of the model given in Section 6, one market maker trades at a high frequency (and is referred to as a 'fast' market maker, with less capital and capacity to bear risk), while the other trades at a low frequency (ie. is 'slow', with higher capital levels and risk bearing capacity). Both market makers set their bid/ask prices and manage their inventories in a way that manages both the risk of bearing losses on their inventories due to changes in market prices (i.e. their market risk) (see (Ho and Stoll (1981)), and the risk of trading with a counterparty who has better information as to the fundamental value of the security (i.e their risk of adverse selection) (Glosten and Milgrom (1985)).

(ii) A fundamental trader holds a belief as to the price of a security that is in line with economic fundamentals. They submit market orders, buying (selling) more when market prices are lower (higher) than their estimate of fundamental value. Although the fundamental trader trades using market orders, it contributes to the resilience and liquidity of the market, insofar as its demand for securities counters any movement in price away from their estimate of fundamental value (see Harris (2003)).

(iii) A momentum trader trades frequently using market orders, demanding liquidity by seeking to trade in the direction of recent price moves. They could be thought of as representing agents whose behaviour is highly sensitive to recent changes in price, such as that stemming from directional trading strategies, stop-loss orders and options hedging strategies (see Section 2).

Like the fast market maker, in examining flash episodes in Section 6 we also increase the frequency with which the momentum trader can trade.

(iv) A noise trader demands liquidity by submitting market orders. They buy and sell with equal probability and in random (normally distributed) size. The noise trader represents the aggregate shock to supply/demand that might enter markets for a variety of different reasons in the real world, the rationale for which lies outside the scope of this model (see Kirilenko et. al (2017)).

[5] Menkveld (2013) finds that traders exhibiting market making strategies overwhelmingly trade with limit orders.

For brevity, we refer to these agents as being of types MM_{slow}, MM_{fast}, *fun*, *mom* and *noise* in the equations that follow.

These types of traders are consistent with the categorisation of traders described elsewhere in the literature (in particular, see Harris (2003)). They are also common in the agent-based model literature that attempts to replicate stylised facts from financial returns (Franke and Westerhoff (2012)), as well as in a high-frequency setting (see Leal et al (2016)).

Agents differ in the average frequency with which they can trade. This is incorporated stochastically into the model – that is, each trader has an average frequency with which it will be allowed to trade, and their actual frequency of trading varies randomly around this.[6] Specifically, in every time period *t*, each agent of type $k \in \{MM_{slow}, MM_{fast}, noise, fun, mom\}$ is able to submit a trade if a random binary, variable, $\Lambda_{k,t}$ takes value 1, where:

$$\Lambda_{k,t} = \begin{cases} 1 \; with \; probability \; \pi_k \\ 0 \; with \; probability \; 1 - \pi_k \end{cases} \qquad (1)$$

where π_k is the probability with which an agent of type *k* trades in any period.

Each trading period *t* represents a discrete time period of one second. We set the total numbers of time periods, *T*, in the model to 27,000: the number of seconds in a 7.5 hour trading day. This is a sufficient number of time intervals to allow for the study of flash events, which have been observed to emerge within a matter of minutes.[7]

The remainder of this section describes the heuristics through which each agent trades in more detail. These are also summarised in **Table 1**.

Table 1: Description of the agents in the model

Market participant type	Heuristic/trading strategy	Order type
Fundamental trader	Buys (sells) if market prices are lower (higher) than its estimate of fundamental value.	Market orders
Momentum trader	Buys (sells) if prices have been rising (falling).	Market orders
Noise trader	Trades with a random size and in a random direction.	Market orders
Slow market-maker	Earns profits from buying at the bid price and selling at the ask. Aims to manage risk associated with holding its inventory and avoid adverse selection. Has higher initial capital levels than the competing, 'fast', market maker.	Limit orders
Fast market-maker	Earns profits from buying at the bid price and selling at the ask. Aims to manage risk associated with holding its inventory and avoid adverse selection. Has lower initial capital levels than competing, 'slow', market maker.	Limit orders

Throughout, we define the demand for liquidity of an agent of type *k* at time *t* as $d_{k,t}$. We adopt the convention that *d* takes positive values for buy and negative values for sell orders when $k \in \{noise, fun, mom\}$. For the supply of liquidity (limit orders) by market makers, we adopt the notation ${q_{k,t}}^{ask}$ where superscript denotes the direction of the limit order (ask or bid), when $k \in \{MM_{slow}, MM_{fast}\}$.

[6] This specification is based on that in Booth (2016).

[7] In principle the frequency of trading of agents in the model could be increased beyond this, but the ensuing computational burden prevents us from doing so here.

(i) Noise trader

The **noise trader** posts market orders of a random size and direction (ie. buy/sell). They do not rely on any notion of fundamental price. We allow them to place buy/sell orders with equal probability, and of a size that is normally distributed, with a mean of zero. That is, we assume the noise trader to have a demand function:

$$d_{noise,t}(\Lambda_{noise,t}, \varepsilon_t) = \Lambda_{noise,t} . (\varepsilon_t) \qquad (2)$$

$where$: $\varepsilon_t \sim N(0, \sigma_{noise})$, for some positive constant σ_{noise}.

(ii) *The fundamental trader*

The **fundamental trader** submits market orders to buy when the observed best-ask price in the central limit order book $ask^*_{(t-1)} = \min(\boldsymbol{p}^{\boldsymbol{ask}}_{\boldsymbol{t-1}})$ is low relative to its estimate of the fundamental value, and sells when the observed best-bid price $bid^*_{(t-1)} = \max(\boldsymbol{p}^{\boldsymbol{bid}}_{\boldsymbol{t-1}})$ it is high relative to fundamentals. The larger the deviation in the observed mid-price from the fundamental price, the more it is incentivised to buy or sell, and the larger order it will post.

The demand function for the fundamental trader is therefore set to:

$$d_{fun,t}(\Lambda_{fun,t}, \omega_{fun}, \vartheta_t, ask^*_{t-1}, bid^*_{t-1}) = \begin{cases} \Lambda_{fun,t}\left[\omega_{fun}.(\vartheta - ask^*_{t-1})\right] if\ \vartheta > ask^*_t\ ; (buy\ order) \\ \Lambda_{fun,t}\left[\omega_{fun}.(\vartheta - bid^*_{t-1})\right] if\ \vartheta < bid^*_t\ ; (sell\ order) \\ 0\ if\ bid^*_t < \vartheta < ask^*_t\ ; \qquad Do\ nothing \end{cases} \qquad (3)$$

where ϑ_t is the fundamental trader's estimate of the fundamental value of the security, and $\omega_{fun} > 0$ is a parameter controlling the sensitivity of their demand to deviations from fundamental value.

The fundamental trader's estimate of the fundamental price follows a random walk. This reflects the continuous flow of information on the fundamental value of the security. Specifically it evolves according to:

$$\vartheta_t = \vartheta_{t-1} + \mu_t\ ;\ where\ \mu_t \sim N(0, \sigma_v) \qquad (4)$$

for some positive constant σ_v.

(iii) *The momentum trader*

As is common in the agent-based modelling literature (see Franke and Westerhoff (2012)), we incorporate a momentum trader (sometimes referred to as a 'chartist'). The **momentum trader** trades in the direction of recent movements in the mid-price (defined as: $\bar{p}_t = \frac{ask^*_{(t)} + bid^*_{(t)}}{2}$) movements by buying (selling) when prices have been rising (falling). The size of its orders increases with the magnitude of recent changes in the price of the asset. Such momentum trading is, in the real world, unlikely to be infinite in scope, however. Reflecting this, the momentum trader also has an inventory limit – that is, a cap to how far it can accumulate net positions by buying or selling. As the momentum trader's position I_{mom} approaches this limit, it gradually reduces the size of its orders (see Anderson and Noss (2013)).

We set the demand function for the momentum trader to

$$d_{mom,t}(\Lambda_{mom,t}, \omega_{mom}, \bar{I}_{mom}, \bar{p}_{t-1}, \bar{p}_{t-l_{mom}}, h)$$

$$= \Lambda_{mom,t} \cdot \left[\underbrace{\omega_{mom}}_{\substack{sensitivity \\ to\ price\ movements}} \cdot \underbrace{\frac{(\bar{p}_{t-1} - \bar{p}_{t-l_{mom}})}{\bar{p}_{t-l_{mom}}}}_{\substack{price\ change\ observed \\ between\ t-1\ and\ t-l_{mom}}} \cdot \underbrace{\left(1 - \left(\frac{l_{mom,t-1}}{\bar{l}_{mom}}\right)^{h}\right)}_{\substack{reduction\ for \\ position\ limit}} \right] \quad (5)$$

where ω_{mom} governs the sensitivity of order size to price movement, l_{mom} governs the period over which price changes are observed, and h controls the amount by which the momentum trader reduces the size of its orders as it approaches its position limit.

(iv) ***The market makers***

The **market makers** $k \in \{MM_{slow}, MM_{fast}\}$ are the most complex agents in the model. Market makers post limit orders against which other participants can trade. Menkveld (2013) finds that eighty per cent of high-frequency market makers' orders are passive, motivating our simplification that they only post limit orders. This does not preclude them from consuming liquidity. In return for providing immediacy, they attempt to 'earn the bid-ask spread' by buying at the price at which they submit limit orders to buy, and selling them at those at which they submit limit orders to sell. The difference – or 'spread' – between the price at which they submit limit orders to buy and sell compensates the market maker for the possibility of them making a loss on their inventory of securities, or for the risk of being adversely selected. The market maker also faces a limit on the quantity of securities they are willing to hold in the hope of making profits, due to a constraint on their available capital (Kirilenko et. al (2017)).

In each period the size and direction of market makers' trading depends on a number of variables:

- *The total volume of limit orders (or quantity of liquidity) it offers at time t*

  The total quantity, $Q_{\mathrm{k,t}}$ of liquidity market makers are willing to supply is a fixed proportion of their available capital $K_{k,t}$ each period:

$$Q_{k,t}(\omega_k, K_{k,t}) = \omega_k, K_{k,t}. \quad (6)$$

  The capital available to the market makers, $K_{k,t}$, is equal to the market maker's starting level of capital $K_{k,0}$, plus the cumulative profits earned from netting trades to buy and sell up to time t, and the mark-to-market profit or loss from its remaining inventory holdings:

$$K_{k,t} = \underbrace{K_{k,0}}_{Initial\ capital} + \underbrace{min\left(\sum_{m_{k,t-1}} d_i^{bid}, -\sum_{m_{k,t-1}} d_i^{ask}\right)}_{\substack{Units\ of\ asset\ traded\ which\ result\ in\ no\ held\ inventory - \\ ie\ those\ where\ buy\ and\ sell\ trades\ are\ netted}} \cdot \underbrace{\left(\frac{\sum_{m_{k,t-1}} p_i^{ask}.d_i^{ask}}{\sum_{m_{k,t-1}} d_i^{ask}} - \frac{\sum_{m_{k,t-1}} p_i^{bid}.d_i^{bid}}{\sum_{m_{k,t-1}} d_i^{bid}}\right)}_{\substack{Weighted\ average\ sell\ price\ minus \\ weighted\ average\ buy\ price}}$$

$$+ \underbrace{\left[max\left(\sum_{m_{k,t-1}} d_i^{bid}, -\sum_{m_{k,t-1}} d_i^{ask}\right) - min\left(\sum_{m_{k,t-1}} d_i^{bid}, -\sum_{m_{k,t-1}} d_i^{ask}\right)\right].\bar{p}_{t-1}}_{\substack{Remaining\ inventory \\ evaluated\ at\ observed\ market\ mid-price}} \quad (7)$$

  …where $m_{k,t}$ is the history of all trades to buy and sell, transacted by participant *k*, up to time *t.*

- *The balance between limit orders to buy (bid liquidity) versus those to sell (ask liquidity)*

Equations (8) and (9) govern the quantity of limit orders to buy, $q_{k,t}{}^{bid}$, and sell $q_{k,t}{}^{ask}$ placed by the market maker. Market makers first split the total quantity of liquidity they are willing to supply equally between orders to buy and sell. This is then sub-divided into *n* distinct limit orders placed on each side of the order book. The market maker adjusts from this equal split in order to manage their inventory relative to a desired level. If it has accumulated a long position in the asset it scales down the quantity of bid orders it posts relative to the quantity of ask orders it posts, and vice versa.

The market maker makes a further adjustment to account for the risk of adverse selection (see Benos and Wetherilt (2012)). That is, they scale down the number of orders they supply on both sides as the size of recent price movements increases.

Bringing this together:

$$q_{k,t}{}^{bid}\left(\Lambda_{k,t}, Q_{k,t}, n, I_t, \bar{I}_k, r, \bar{p}_{t-1}, \bar{p}_{t-l_k}\right) = \underbrace{\Lambda_k}_{\substack{1 \text{ if they can submit,} \\ 0 \text{ otherwise}}} \cdot \underbrace{\frac{1}{2}Q_{k,t}}_{\substack{Half\ of\ total\ liquidity \\ they\ wish\ to\ post}} \cdot \underbrace{\frac{1}{n}}_{\substack{Split\ orders \\ into\ n\ levels}} \cdot \underbrace{\left(\frac{I_t - \bar{I}_k}{\bar{I}_k}\right)}_{\substack{adjustment\ for \\ inventory\ held}} \cdot \underbrace{r\left(1 - \left|\frac{\bar{p}_{t-1}}{\bar{p}_{t-l_k}}\right|\right)}_{\substack{Reduction\ for\ fear \\ of\ adverse\ selection}} \quad (8)$$

$$q_{k,t}{}^{ask}\left(\Lambda_{k,t}, Q_{k,t}, n, I_t, \bar{I}_k, r, \bar{p}_{t-1}, \bar{p}_{t-l_k}\right) = \underbrace{\Lambda_k}_{\substack{1 \text{ if they can submit,} \\ 0 \text{ otherwise}}} \cdot \underbrace{\frac{1}{2}Q_{k,t}}_{\substack{Half\ of\ total\ liquidity \\ they\ wish\ to\ post}} \cdot \underbrace{\frac{1}{n}}_{\substack{Split\ orders \\ into\ n\ levels}} \cdot \underbrace{\left(\frac{I_t + \bar{I}_k}{\bar{I}_k}\right)}_{\substack{adjustment\ for \\ held\ inventory}} \cdot \underbrace{r\left(1 - \left|\frac{\bar{p}_{t-1}}{\bar{p}_{t-l_k}}\right|\right)}_{\substack{Reduction\ for\ fear \\ of\ adverse\ selection}} \quad (9)$$

where $I_{k,t}$ is the market makers' inventory at time *t*, and $\bar{I}_k$ is its maximum inventory level. The strength of the adverse selection heuristic is governed by r, and l_k governs the period over which price changes are observed.

- *The market makers' choice of mid-price*

The market maker changes its prices according to a mixture of two heuristics:

1. The first is related to the volume of incoming market orders it observes during a given trading period. The larger demand to buy (sell), the greater the degree to which the market maker will infer a signal that the fundamental price of the asset is higher (lower) than its previous level, and increase (decrease) its prices accordingly.

   We define the market wide demand D_t as the demand from all the traders submitting market orders in period *t*. We can then define a demand signal observed by the market maker $\widetilde{\mathrm{D}}_{\mathrm{t}}$ as:

$$\widetilde{D}_t = \begin{cases} D_1 \ if & t = 1 \\ \alpha(D_t) + (1-\alpha)\widetilde{D}_{t-1} & t > 1 \end{cases} \quad (10)$$

   where α is a weighting parameter such that $0 < \alpha < 1$.[8] The higher the value of α, the greater the extent to which the market maker discounts demand in previous periods, and conditions its behaviour to a greater degree on the more recent past.

[8] In this set up the market makers' behaviour depends only on demand that it observes in the form of market orders (aggressive orders) placed by other (liquidity demanding) types of participant. This is for computational simplicity, and rules

2. The market maker uses prices as a way to manage the market risk associated with its inventory I_t of securities (see Ho and Stoll (1981)). It therefore adjusts its mid-price up (down) if it is short (long), to discourage additional sales/purchases that would otherwise expand its inventory. We therefore define $\tilde{I}_{k,t}$, as:

$$\tilde{I}_{k,t} = \begin{cases} c.\left(\frac{I_{k,t}}{\bar{I}_k}\right) & if \quad \mathbb{I}[|I_{t-1} - I_{t-2}| > 0] \\ \tilde{I}_{k,t-1} & if \quad \mathbb{I}[I_{t-1} - I_{t-2} = 0] \end{cases} \quad (11)$$

where c is the sensitivity of the mid-price to changes in its inventory, and $\mathbb{I}$ is the indicator function.

The market maker then sets its mid-price by looking at the previous mid-price in the market, and then adjusting it downwards (upwards) as a function of observed net market-wide demand for liquidity on the buy-side (sell-side) beyond a constant threshold $\underline{s}$, and adjusting it downwards (upwards) if it in the previous period had to increase its inventory.

That is, its mid-price $\overline{p^*}_{t,k}$ is set so that:

$$\overline{p^*}_{t,k}(\tilde{D}_t, \tilde{I}_{k,t}) = \underbrace{\overline{p^*}_{t-1}}_{\substack{Previous\ market \\ wide\ mid-price}} + \underbrace{\mathbb{I}[\tilde{D}_t > \underline{s}].\tilde{D}_t}_{\substack{Price\ impact\ of \\ observed\ buying \\ from\ market\ orders}} - \underbrace{\mathbb{I}[\tilde{D}_t < -\underline{s}].\tilde{D}_t}_{\substack{Price\ impact\ of \\ observed\ selling \\ from\ market\ orders}} + \underbrace{\tilde{I}_{k,t}}_{\substack{Adjustment\ for \\ inventory \\ position\ if\ inventory\ changed}} \quad (12)$$

- *The bid-ask spread it requires around the mid-price*

The choice of prices at which the market maker submits limit orders around this mid-price increases with its susceptibility to adverse selection (as in Ho and Stoll (1981), who provide a theoretical model which results in a bid-ask spread that correlates positively with market volatility). We proxy this risk of adverse selection by the volatility of mid prices in the previous ten periods ($\sigma_{t-1:t-10}$), with market makers setting their bid-ask spread as a linear increasing function of this volatility, subject to a constant minimum, ξ.

The market makers' *i*th limit order to sell and buy are then set at prices $\mathrm{p}_{\mathrm{k,t}}^{\mathrm{ask,i}}$ and $\mathrm{p}_{\mathrm{k,t}}^{\mathrm{bid,i}}$ where:

$$p_{k,t}^{ask,i}(\bar{p}_{t-1}, \delta_t, \alpha, \gamma, \sigma_t) = \overline{p^*}_{t,k} + \min(\gamma\sigma_{t-1:t-10}, \xi) + i.(10^{-\tau}) \quad (13)$$
$$p_{k,t}^{bid,i}(\bar{p}_{t-1}, \delta_t, \alpha, \gamma, \sigma_t) = \overline{p^*}_{t,k} - \min(\gamma\sigma_{t-1:t-10}, \xi) - i.(10^{-\tau}) \quad (14)$$

where $\overline{\mathrm{p}^*}_{\mathrm{t,k}}$ is the mid-price in the market at time *t* and i governs the sensitivity of the bid-ask spread to volatility. The market makers set best bid and ask prices, and set further orders above and below these prices, totalling n orders on each side of the order book. The parameter τ governs how far apart these orders are spread (ie. it corresponds to the 'tick size').

In summary, therefore, the model consists of a financial market consisting of

- A noise trader that posts market orders of a random size, $d_{noise,t}$.

---

out the possibility that one market maker responds to limit orders placed by another market maker – which otherwise greatly increases the complexity of the model and its solution.

- A momentum trader that posts market orders of size $d_{mom,t}$, with the aim of trading in the direction of recent price moves, subject to a cap on its inventory.

- Market makers that post limit orders to buy and sell:

    - Of size ${q_{k,t}}^{bid}$ and ${q_{k,t}}^{ask}$, such that their total supply of liquidity is proportional to their available capital.

    - At a bid and ask prices $p_{k,t}^{bid}$ and $p_{k,t}^{ask}$ that are based on its inference as to the fundamental price and need to manage the market risk associated with their inventory;

      and

    - At a bid-ask spread a spread designed to compensate it for risk.

Thus the model is fully characterised by the set of strategies: $\{d_{noise,t}, d_{mom,t}, d_{mom,t}, {q_{k,t}}^{bid}, {q_{k,t}}^{ask}, p_{k,t}^{ask}, p_{k,t}^{ask}\}$, initial values for prices and capital, and a market clearing mechanism (the central limit order book).

4. Calibration

We calibrate the model by choosing values for five of its parameters so that the dynamics of simulated prices match those observed empirically. For this baseline specification, we assume that all market participants – including the fast market maker and momentum traders – trade with the same frequency.

These five parameters whose value we calibrate are:

- The frequency with which market makers are able to trade, π_{MM} , and the weight they place on past observed demand by other participants in determining their own demand function, α_{MM}.

- The frequency with which momentum traders are able to trade, π_{mom}, their demand to changes in prices ω_{fun}.

- The frequency with which fundamental traders are able to trade, π_{fun}.

All other parameters take fixed values, the details of which are given in **Annex 1**.

Data used in the calibration are a time series of high-frequency price data provided by Thomson Reuters on the sterling/US-dollar exchange rate over the period 3 - 7 October 2016. Data are taken from the Thomson Reuters Matching platform – a central limit order book that supports both market and limit order types. The Thomson Reuters Matching platform is thought to account for around 5-10% of trading in the sterling spot market in normal conditions (see Noss et al (2017)).

Estimated parameter values are those that give rise to modelled prices whose moments match those estimated empirically. The moments considered here follow those considered elsewhere in the literature (in particular see Cont (2001), which considers certain stylised facts of financial prices; and Gould et al (2013) which does so in the context of models of a Limit Order Book). These moments are:

(i) The volatility of the mid-price

We use the mean absolute returns as our estimate of volatility, following Cont (2001).

(ii) The degree to which the distribution of returns is heavy tailed

The distribution of changes in financial market prices have been shown to exhibit heavy tails at almost all time scales (Gould et al 2013). Following Cont (2001), we use the Hill estimator of the tail index to estimate the degree of heavy-tailedness in the distribution of returns on the mid-price. A lower value implies this distribution has fatter tails.

(iii) The autocorrelation of returns

Except for weak negative correlation at extremely short timescales, autocorrelations in financial price time series have generally been found to be insignificantly different from zero (Bouchaud and Potters (2003) document this in the case of FX markets).

We therefore calibrate the model to an autocorrelation of zero over windows of both 60 lags and 900 lags (corresponding to time windows of 1 minute and 15 minute, respectively). This allows us to verify this stylised fact at both higher and lower frequencies.

(iv) The autocorrelation of the volatility of the mid-price (volatility clustering)

Absolute mid-price returns have been documented to display long memory at intraday timescales. That is, volatility clusters over time, with a period of higher volatility more likely to be followed by a period of high volatility.[9] Since in our model the fundamental price evolves according to a random walk, it is the interactions of the agents that allows the model to replicate this property of volatility clustering.

To capture this we include the estimated autocorrelation coefficients over an arbitrarily chosen set of lags covering frequencies 1, 60, 300 and 900 periods.

The estimated empirical values for these moments and confidence bands around them are given in **Table 2**. These empirical confidence intervals are calculated asymptotically using the empirically observed mid-price on 3 October 2016.

The five parameters are estimated by matching the model outputs to these empirical moments. This is achieved by maximising the 'joint coverage ratio' – a technique proposed in Franke and Westerhoff (2012) (and used to calibrate the agent-based model in Braun-Munziger et al (2016)). This criterion finds parameter values that maximise the proportion of simulated price series for which the above moments of the simulated series of prices fall within 98% confidence intervals of their empirical counterparts. This is achieved by means of a numerical grid search over a feasible bounded set of parameters.[10]

The estimated parameters are shown in Table 3 and their respective moment coverage ratios are shown in Table 4.

Figure 1 shows the distributions of the moments of 1000 simulation runs, as well as their empirical moments and confidence intervals. The joint moment coverage ratio (Table 4) indicates that the probability of all the moments of a given model simulation lying within the 98% confidence interval of their empirical value is around 50%.

[9] Cont et al (1997) document this for the USD-JPY exchange rate. This property has been found to apply to returns on other securities including futures and equities (Zhao (2010)).

[10] Whilst this may not ensure a globally optimal solution, it is necessary in order to keep the computational burden manageable.

Table 2 – Empirical Moments and confidence bands				
Moment	**Notation**	**Lower confidence band**	**Sample mean**	**Upper confidence band**
(i) Mean absolute mid-price returns	ϑ	0.0007	0.0008	0.0009
(ii) Hill estimate (inverse)	$\frac{1}{\alpha}$	0.3428	0.3851	0.4274
(iii) Autocorrelation of returns (60 lags)	ρ_{60}^{r}	-0.0253	0.0015	0.0282
(iii) Autocorrelation of returns (900 lags)	ρ_{900}^{r}	-0.0159	0.0084	0.0327
Volatility autocorrelation (1 lag)	ρ_{1}^{σ}	0.1985	0.2346	0.2708
Volatility autocorrelation (60 lags)	ρ_{60}^{σ}	0.0027	0.0357	0.0686
Volatility autocorrelation (300 lags)	ρ_{300}^{σ}	0.0014	0.0296	0.0579
Volatility autocorrelation (900 lags)	ρ_{900}^{σ}	-0.0208	0.0043	0.0294

Table 3 – Estimate parameter values (baseline)	
Parameter	**Estimated value**
Market maker speed	0.28
Signal decay	0.15
Momentum traders speed	0.03
Fundamental traders speed	0.15
Momentum trader sensitivity	100000

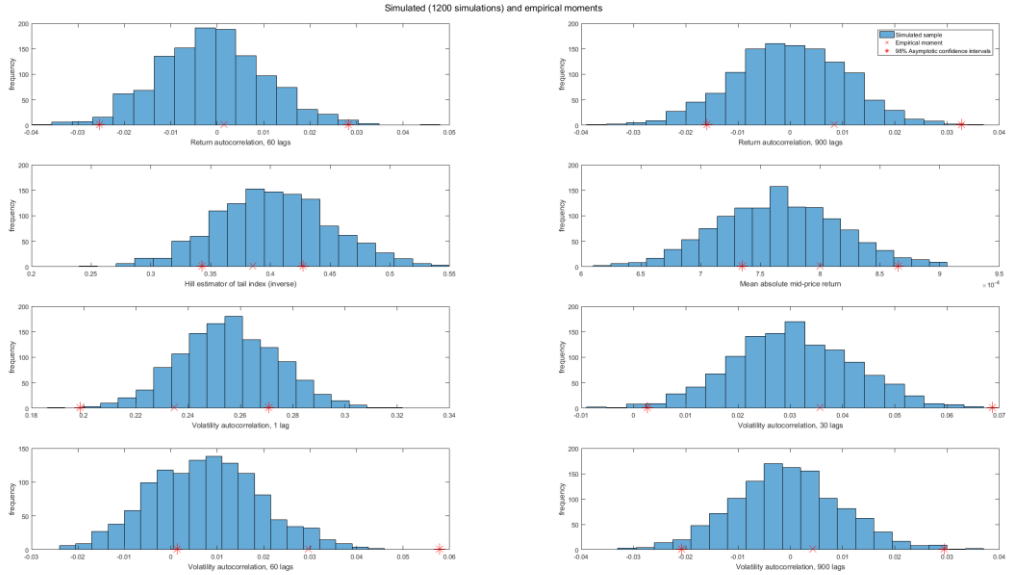
Table 4 – Moment coverage ratios in the baseline model	
Moment	**Moment coverage ratio**
Return autocorrelation (60 lags)	100%
Return autocorrelation (900 lags)	97%
Hill estimator of tail index (inverse)	70%
Mean absolute mid-price returns	80%
Volatility autocorrelation (1 lag)	97%
Volatility autocorrelation (60 lags)	97%
Volatility autocorrelation (300 lags)	80%
Volatility autocorrelation (900 lags)	95%
Joint moment coverage ratio	53%

As a cross-check on the validity of our calibration strategy, we also estimate the model using the 'method of simulated moments', described in Franke and Westerhoff (2012). Under this methodology, the optimised parameter set is that which minimises a function of the difference between the moments of the simulated series of prices, $\boldsymbol{m}^{sim}$ and their empirical counterparts $\boldsymbol{m}^{emp}$. We calculate two standard method of simulated moments loss functions – one using an Ordinary Least Squares loss function $J^{OLS} = (\boldsymbol{m}^{sim} - \boldsymbol{m}^{emp})\boldsymbol{I}(\boldsymbol{m}^{sim} - \boldsymbol{m}^{emp})'$, where **I** is the identity matrix. We also calculate the asymptotically efficient estimator: the Weighted Least Squares function $J^{WLS} = (\boldsymbol{m}^{sim} - \boldsymbol{m}^{emp})\boldsymbol{W}(\boldsymbol{m}^{sim} - \boldsymbol{m}^{emp})'$, where the Weighting matrix W is an estimate of the inverse of the covariance matrix $\widehat{\sum}^{-1}$ across the simulated moments. This is estimated using a block bootstrap method on the empirical data.

Across parameter sets and noise paths, both the mean loss functions J^{OLS} and J^{WLS} correlate strongly and negatively with the joint moment coverage ratio, with statistically significant correlation coefficients of -0.64 and -0.60 respectively. Minimising both J^{OLS} and J^{WLS} would result in the selection of a different parameter set than the joint moment coverage ratio, although both solutions are contained in the 3rd percentile of their sample distributions. That provides further validation that the selected parameter set is appropriate.

We conclude that – conditional at least on the appropriateness both of our choice of moments and values of imposed parameters – this calibration provides a sensible basis on which to proceed.

Figure 1 – Simulated and empirical moments



5. Simulation results (benign market conditions)

Figure 2 compares empirical time series of mid-prices for 3-5 October (blue lines) to three simulated series of mid-prices from the baseline model described in Section 4. At least at sight, the baseline model produces time series of prices whose dynamics appear similar in form to

Figure 2 – Market and simulated mid-prices

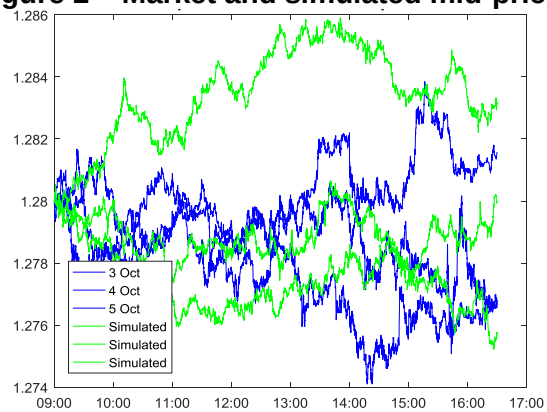
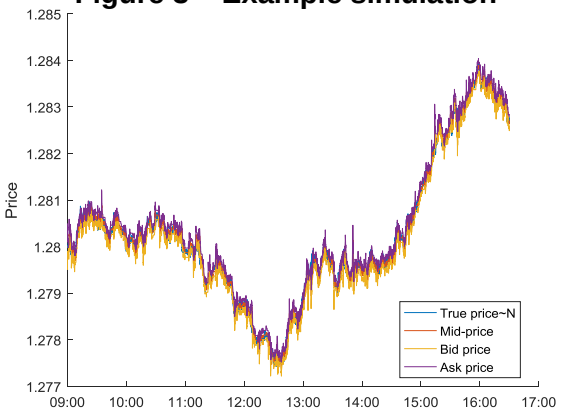


Figure 3 – Example simulation



those observed in the data.

Figure 3 shows a single time series of simulated mid, bid and ask prices. Such single simulations are useful in that they allow us to investigate the behaviour of the baseline model, where all agents observe – and have the opportunity to trade in – the market at the same frequency.

Figures 4 and 5 show how the positions of each trader evolve over the course of a single simulation, due to their net buying behaviour. From this we can observe how the different types of market participants interact to ensure the stability and continuity of market prices:

Figure 4 – Market maker inventory

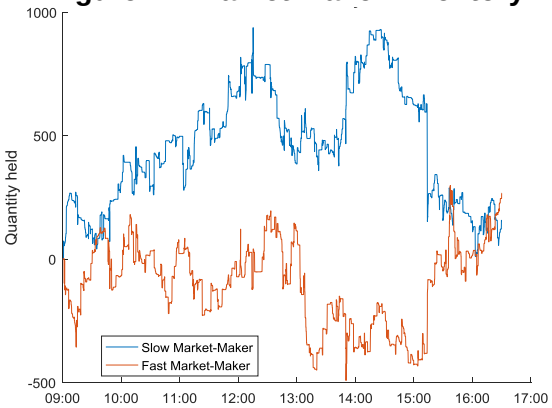


Figure 5 – Other traders' inventory

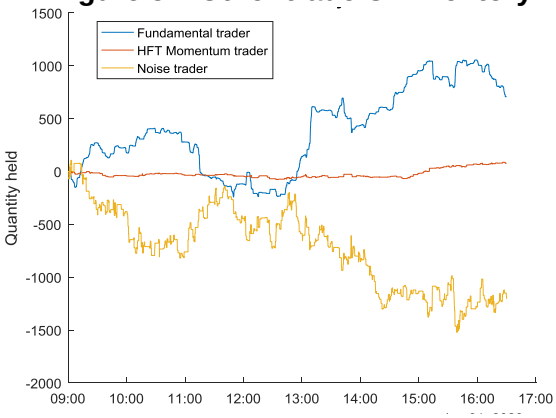


Figure 6 – Market maker price pressure

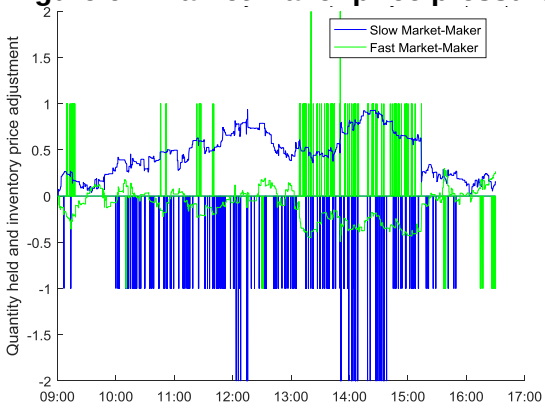
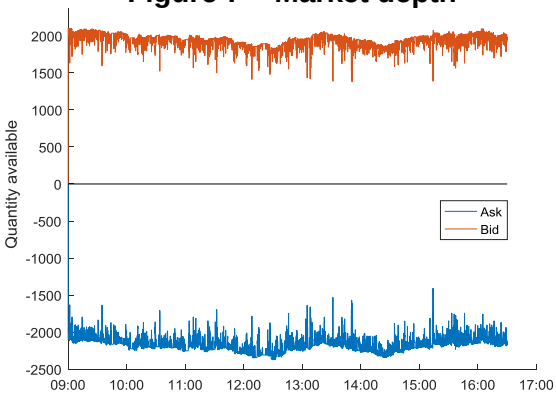


Figure 7 – Market depth



- Although the noise trader buys and sells with equal probability, over the course of this particular simulation it happens to accumulate a net short position (yellow line in Figure 5).

- Between 12:00 and 13:00, relatively high demand to sell by the noise trader results in downward pressure on the mid-price (Figure 3).

- Market makers perceive this selling as a signal to lower their mid-price. These adjustments are shown by the blue bars in Figure 6.

- Both of these effects induce buy orders from the fundamental trader (blue line in Figure 5) – which returns the price to a level closer to that commensurate with fundamentals.

- The momentum trader has, as a result of low levels of price movement, remained fairly inactive (red line in Figure 5), but accrues a net long position during the rapid increase in price that occurs from around 15:00.

Figure 7 shows the market depth on each side of the order book (the sum of the quantities of all limit orders). Order book depth falls during the most volatile parts of the day, such as at 13:00 on the bid side and 15:15 on the ask side, but remains relatively resilient to the increased price volatility.

In summary, when – as is the case of this baseline model specification – all agents have the opportunity to trade at the same frequency, this gives rise to relatively stable market prices whose changes are relatively continuous. Initial selling pressure from the noise trader is reinforced by the withdrawal of liquidity by the market markers, but any resulting move in price is arrested by the stabilising actions of the fundamental trader, which restores prices to their estimate of equilibrium value.

6. Flash episodes

We also use the model to investigate market dynamics during flash episodes, and how these may be driven by the presence of high-frequency trading.

To introduce high-frequency traders into the model, we increase the probability with which the fast – but not the slow – market-maker and momentum trader can trade, compared to that in the baseline specification. This matches the intuition that these 'fast' market participants are high frequency traders that, on average, trade more often than their low frequency counterparts.

The resulting time series of prices contain flash episodes – that is, large movements in market mid prices away from fundamentals that quickly reverse. Figures 8 shows such a single simulation of prices. Analogous to the analysis in the previous section, examining such a single simulation gives some insight into how the trading behaviour of different market participants interact to give rise to the flash episode, and how they do so in a way that roughly matches the drivers of *real-world* episodes discussed in Section 2:

- A large initial sell order from the (low frequency) noise trader is executed just prior to 14:13:30 (yellow line in Figure 10). This has the effect of depleting available volume of limit orders to buy, lowering the mid-price.

- In response to the incoming market order from the noise trader, the high-frequency market maker infers that the fundamental price of the asset is lower than its current level. It therefore lowers its mid-price.

- A number of procyclical behaviours result amongst the high-frequency traders. These reinforce the resulting movement in price:

    - First, procyclical *liquidity demand* arises from the momentum trader (who is operating at high-frequency), who sells the asset in response to the fall in mid-price (red line in Figure 10).

    - Second, there is a simultaneous rapid *withdrawal* of limit orders (red line in Figure 10) as the fast market maker withdraws their provision of liquidity in response to the incoming market orders from the noise and momentum traders, in order to avoid the risk of being adversely selected. Market depth is severely reduced (spikes in Figure 11); and the market maker widens their bid-ask spread in response to the volatility of the mid-price (Figure 8).

      These procyclical behaviours roughly match those that have exacerbated flash episodes observed in reality (see Section 2). In addition, the behaviour of the momentum trader might match that of investors with directional trading strategies seen during the 2010 equity flash crash, or the behaviour of dealers in hedging their positions and executing client orders in the 2016 sterling flash episode.

- The withdrawal of liquidity by the fast market maker means that the market orders placed by the momentum trader are executed against limit orders of the slow market maker. The inventory of the slow market maker therefore expands passively (blue line in Figure 9).

Figure 8 – Example of a flash event

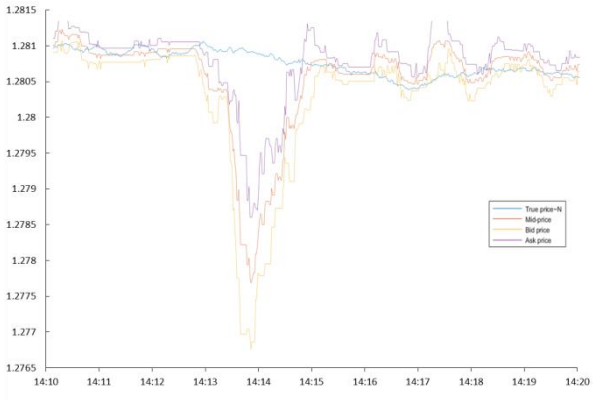


Figure 9 – Market maker inventory

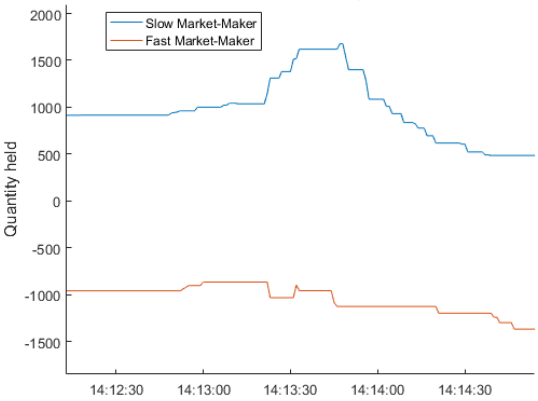


Figure 10 – Traders' inventory

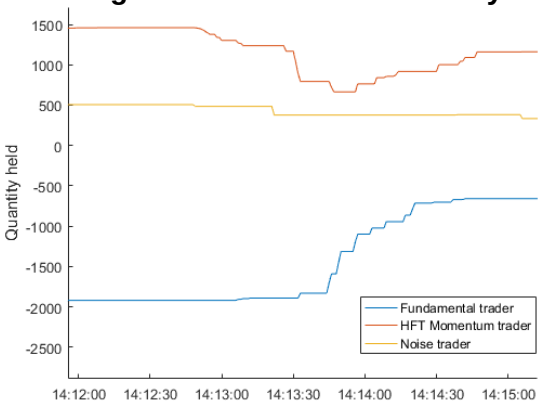
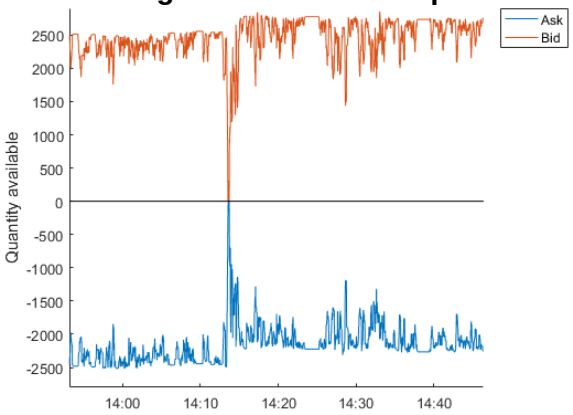


Figure 11 – Market depth



This sharp reduction in the supply of, and increase in demand for, liquidity, rapidly becomes self-reinforcing. The withdrawal of liquidity by the high-frequency market maker causes further falls in the mid-price. The momentum trader takes this as a signal to sell, demanding further liquidity, which – in the face of reduced liquidity supply – results in a sharper movement in price. This in turn further reduces the supply of liquidity by the fast market maker.

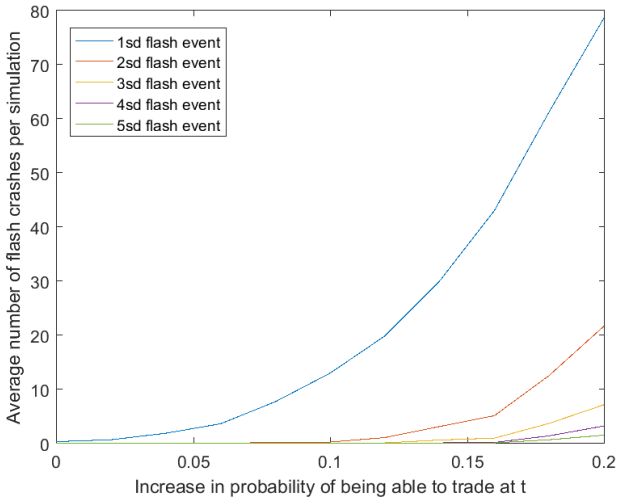
This rapid and self-fulfilling fall in price is only reversed when the low-frequency fundamental trader finally enters the market, and responds to the fall in price away from the level commensurate with fundamentals (Figure 8). In doing so, it places orders to buy that stabilise prices, and in doing so expands its inventory (Figure 10).

In summary, the flash episode arises from markets becoming temporarily dominated by the procyclical behaviours of high-frequency participants. This is triggered by a random order from the noise trader that is large relative to the supply of available limit orders. This then leads to a reduction in the supply of liquidity by market makers operating at high frequency, and a reduction in their mid-price, in order to avoid the risk of their being adversely selected by a market participant with more up-to-date information on the fundamental value of the asset. These dynamics take hold over a time horizon shorter than that over which lower-frequency participants observe the market. The resulting cycle of increasingly depleted liquidity and sharp movement in prices is only reversed once the slower moving fundamental trader re-enters play.

How the frequency of flash crashes changes with the relative frequency with which low- and high-frequency traders participate

As described in Section 2, the recent occurrence of flash crashes has occurred in markets with a high degree of high-frequency trading. It remains uncertain, however, whether the increase in high-frequency trading has precipitated the increased occurrence of flash crashes. We therefore use the model to investigate how the occurrence of flash crashes changes with the prevalence of high-frequency traders.

Figure 12 – Prevalence of flash episodes versus the level of participation of high-frequency traders



To do so, we increase the frequency with which the momentum trader and fast market maker trade, and count the occurrence of flash episodes – which, in this case, we classify as a *k* standard deviation move in price, which reverses over a horizon of less than sixty periods (analogous to a period of a minute in the baseline calibration).[11] The results are shown in Figure 12, which compares the probability of high-frequency market participants (the high-frequency market maker and momentum trader) trading – compared to that in the baseline model – to the average number of flash episodes that occur across model simulations. The number of flash episodes increases in the average frequency with which the high-frequency market participants participate in the market.

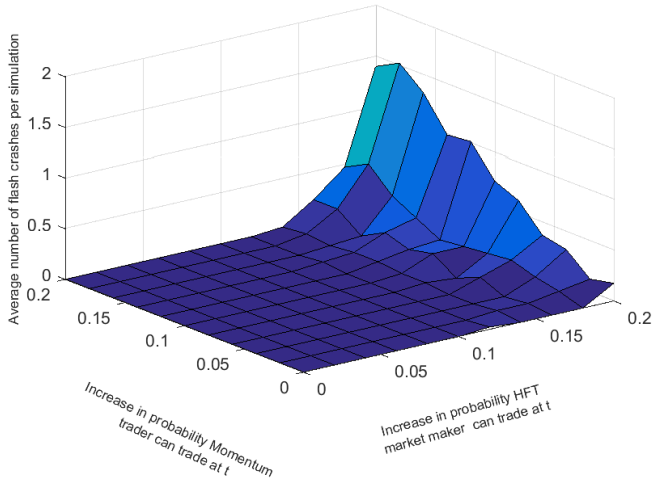
This result fits with the intuition above as to how flash episodes arise. The greater the probability of participation by the higher-frequency momentum and high-frequency market makers, the greater the probability that a sharp (if random) change in price (such as that initiated by the noise trader in the example above) might trigger the emergence of procyclical dynamics.

Importantly, a large initial change in price does not – in itself – give rise to a flash episode. Instead, flash episodes only develop when high frequency traders have a probability of trading that is high enough to allow their procyclical dynamics to take hold over a horizon shorter than that at which slower moving market participants can intervene to restore prices to fundamentals. The greater the *relative* frequency with which the low frequency market participants participate the greater the likelihood that they step in to arrest the resulting falls in price.

We also compare the *relative* effects on the prevalence of large (five standard deviation) flash episodes of trading by the high-frequency momentum trader and the high-frequency market maker participate. This is illustrated in Figure 13. This shows that the participation of both types of market participant – ie. the momentum trader who demands liquidity at high frequency, and the market maker that provides it – is necessary in order to precipitate flash episodes. The participation of one type of high frequency trader alone is insufficient.

Figure 13 – How the prevalence of flash episodes changes with the participation of the high-frequency market maker and momentum trader

[11] This definition of a flash episode is arbitrary, but is intended to capture how the majority of flash crashes have corrected within a short period.

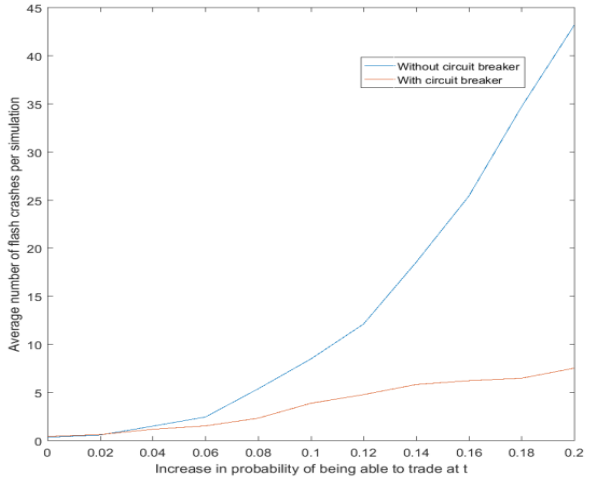


This result confirms the intuition described above that flash episodes occur as a result of the self-reinforcing dynamic that results from the interaction of these two types of high frequency market participant. The withdrawal of the provision of liquidity by the high-frequency market maker leads to a fall in the mid-price of the security, which the momentum trader takes this as a signal to sell, demanding further liquidity, which in turn further reduces the supply of liquidity by the fast market maker, given their fear of adverse selection.

Policy experiments

The model could be used to investigate a number of policy interventions. As an example, we use the model to evaluate the degree to which the introduction of circuit breakers – that is, mandated trading halts that come into effect after a price move of a given magnitude – might curb the frequency and/or severity of flash episodes. Such measures were implemented in the US in response to the 2010 US equity flash crash (see Clapham et al (2018)). We investigate the use of a single instrument trading halt that applies to trading in the single security considered in this model.

Figure 14 – How the prevalence of flash crashes changes with the introduction of a circuit breaker



The calibration and publication of trading halts under MiFID II is an active area of regulatory investigation and rule-making (ESMA (2017)). We introduce a circuit breaker that halts trading for five minutes if the mid-price moves more than 1.5 standard deviations within a period of one minute.[12] In doing so we adapt the model so that, in the period following the end of the trading halt, participants of every type can trade with the same probability for five minutes. This is designed to capture conditions akin to those of a starting auction – a mechanism employed by some exchanges to restart trading after a halt (for further details see FCA (2017) and Ackert (2012)).

Figure 14 shows the prevalence of flash episodes – of a size equal to a one standard deviation change in price – with and without the circuit breaker, and how this varies with the frequency with which the high-frequency traders participate. Mechanically, the introduction of the circuit breaker caps the deviation in market prices beyond 1.5 standard deviations and causes a halt in trading.

[12] Circuit breakers – or mandated trading halts – can be specified in a variety of ways. See ESMA (2017) for a list of key parameters.

The starting auction that follows the trading halt – whereby participants of every type trade with the same probability for a short period – also has the effect of reducing the prevalence of flash crashes, immediately after the trading halt itself. This is why the prevalence of flash episodes of

one standard deviation in size in the presence of the circuit breaker (indicated by the blue line in Figure 14) is – for a given probability of the high-frequency traders' participation – less than that in the absence of trading halts (indicated by the blue line in figure 12). In other words, the opening auction – simulated here via equal participation of all market participant types, regardless of their speed – effectively pools liquidity, mitigating the procyclical dynamics that lead to flash episodes and allowing for the resumption of orderly trading. This mimics the effect of such opening auctions as implemented in some real world trading venues (see Ackert (2012)). The optimal design of such opening auctions, and their effect on market stability, could be the subject of further work.

7. Conclusions

The electronification of financial markets, improvements in technological capabilities and the associated increase in algorithmic trading have been accompanied by flash episodes in major financial markets over the last eight years (Bank of England (2017)).

This paper offers an agent-based model of trading in a single security via a central limit order book with liquidity providing and liquidity consuming participants. This demonstrates how agent behaviours can combine to lead to stable and continuous prices when agents trade with a similar frequency. But such behaviours can nonetheless combine and lead to flash episodes when some agents observe and trade in the market significantly more frequently than others. Such agents increase their demand for liquidity procyclically, whilst others can withdraw their supply of liquidity in response to adverse selection risk. This aligns with some empirical evidence as to the behaviour of market participants during recent market-wide flash episodes. The prevalence of flash episodes increases with the relative frequency with which some agents trade. Under the scope and terms of the model described here, the introduction of mandatory halts in trading (circuit breakers) has the effect of reducing the frequency of flash episodes.

This framework is not without shortcomings. In particular, it focuses on the interaction of agents who differ in the frequency with which they observe (or, analogously, the speed with which they trade in) markets, but whose trading is governed by heuristics that are relatively simple. In common with some other agent based models (see, in particular, Braun-Munzinger et al (2016)), there is therefore no role for more sophisticated decision making by agents, including that arising from a profit maximising objective, the solution to which might vary with market conditions. This has the drawback of meaning that the behaviour of agents might change in the face of certain policy measures (including the application of circuit breakers), meaning that its results might be an unreliable guide to the the efficacy of such policies (a variety of the 'Lucas critique').

Agents' behaviours are also fixed over time. A more complex framework might incorporate changing in styles of trading behaviour. This might include a role for endogenous switching between different investment strategies by agents, of the variety considered by Franke and Westerhoff (2012). Introducing such switching behaviour might reinforce some of the dynamics present in the results, and the perniciousness of flash episodes. For example, as prices decline during stress, some agents might judge it to be profitable to switch from pursuing a strategy based on a notion of market fundamentals, to one based on momentum. This might reinforce the downward trajectory of prices, and delay their recovery.

Such extensions are, however, left as further work.

References

Ackert, L (2012), 'The impact of circuit breakers on market outcomes'. Foresight, UK Government Office for Science.

Anderson, N and Noss, J (2013), 'The Fractal Market Hypothesis and its implications for the stability of financial markets', Bank of England Financial Stability Paper, No. 23.

Anderson, N, Webber, L, Noss, J, Beale, D, Crowley-Reidy, L (2015), 'The resilience of financial market liquidity', Bank of England Financial Stability Paper, No. 34.

Bank of England (2017), 'Financial Stability Report'.

Bank of England (2016), 'Financial Stability Report'.

Bank for International Settlements (2017), 'The sterling 'flash event' of 7 October 2016'.

Bayraktar, E, Munk, A (2017), 'Mini-flash crashes, model risk, and optimal execution' https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2975769.

Benos, E, Brugler, J and Hjalmarsson, E (2015), 'Interactions among high-frequency traders', Bank of England Working Paper, No. 523.

Benos, E and Sagade, S (2012), 'High-frequency trading behaviour and its impact on market quality: evidence from the UK equity market', Bank of England Working Paper, No. 469.

Benos, E and Wetherilt, A (2012), 'The role of designated market makers in the new trading landscape', Bank of England Quarterly Bulletin Q4.

Boehmer, E, Fong K and Wu, J (2015), International evidence on algorithmic trading, Working paper EDHEC Business School.

Bookstaber, R and Paddrik, M (2015), 'An agent-based model for crisis liquidity dynamics', Office of Financial Research Working Paper Series.

Bookstaber, R, Foley, M.D and Tivnan, B (2015), 'Market liquidity and heterogeneity in the investor decision cycle', Office of Financial Research Working Paper Series.

Booth, A (2016), 'Automated Algorithmic Trading: Machine Learning and Agent-based Modelling in Complex Adaptive Financial Markets', University of Southampton.

Braun-Munzinger, K, Liu, Z and Turrell, A (2016), 'An agent-based model of dynamics in corporate bond trading', Bank of England Working Paper, No. 592.

Brewer, P, Citanic, J and Plott, C (2013), 'Market microstructure design and flash crashes: a simulation approach', Journal of Applied Economics.

Cielinska, O, Joseph, A, Shreyas, U, Tanner, J and Vasios, M (2017), 'Gauging market dynamics using trade repository data: the case of the Swiss franc de-pegging', Bank of England Financial Stability Paper No. 41.

Clapham, B, Gomber, P, Panz, S, (2018). Coordination of Circuit Breakers? Volume Migration and Volatility Spillover in Fragmented Markets. SAFE working paper No. 196.

Cont, R (2001), 'Empirical properties of asset returns', Quantitative Finance, Vol. 1, p. 223-236.

Cont, R, Potters, M, Bouchard, JP, Scaling in stock market data: stable laws and beyond. Centre de Physique des Houches book series. Vol 7.

Easley, D and O'Hara, M (1987), 'Price, Trade Size and Information in Securities Markets,' Journal of Financial Economics 19, p. 69-90.

ESMA (2013), 'Calibration of circuit breakers and publication of trading halts under MiFID II', available at: https://www.esma.europa.eu/sites/default/files/library/esma70-872942901-63_mifid_ii_guidelines_on_trading_halts.pdf

FCA (2017), 'Catching a falling knife: an analysis of circuit breakers in UK equity markets'. available at https://www.fca.org.uk/publications/research/analysis-circuit-breakers-uk-equity-markets.

Foucault, T, Hombert, J and Rosu, I (2016), 'News trading and speed', The Journal of Finance, Vol. 71, No. 1 p. 335-382.

Franke, R and Westerhoff, F (2012), 'Structural stochastic volatility in asset pricing dynamics: Estimation and model contest', Journal of Economic Dynamics and Control, Vol. 36, No. 8, p. 1193–1211.

Gao, C and Mizrach, B (2016), 'Market Quality Breakdowns in Equities', Journal of Financial Markets.

Glosten, L. R and Milgrom, P (1985), 'Bid, ask and transaction prices in a specialist market with heterogeneously informed traders', J. L. Kellogg Graduate School of Management Working Paper No. 570.

Gould, M, Porter, M, Williams, S, McDonald, M, Fenn D and Howison, S (2013), 'Limit Order Books', Journal of Quantitative Finance, Vol. 13, p. 1709-1742.

Gomber, P, et al, 'High-Frequency Trading' (2011). Available at SSRN: https://ssrn.com/abstract=1858626

Harris, L (2003), 'Trading and Exchanges', Oxford University Press.

Ho, T and Stoll, H (1981), 'Optimal dealer pricing under transactions and return uncertainty', Journal of Financial Economics, Vol. 9, Issue 1, p. 47-73.

IMF (2015), 'Market liquidity-Resilient or fleeting?', Global Financial Stability Report Chapter two, October. Washington D.C.

Kim, G and Markowitz, H (1989), 'Investment rules, margin, and market volatility', The Journal of Portfolio Management.

Kirilenko, A, Kyle, A, Samadi, M and Tuzun, T (2017), 'The Flash Crash: High-Frequency Trading in an Electronic Market', Journal of Finance, Vol. 72, p. 967-998.

Kyle, A (1985), 'Continuous auctions and insider trading', Journal of the Econometric Society, p. 1315-1335.

Korajczyk, R and Murphy, D (2014), 'High Frequency Market Making to Large Institutional Traders', Working Paper.

Leal, S, Napoletano, M, Roventini, A and Fagiolo G (2014), ‘Rock around the clock: An agent-based model of low- and high-frequency trading’, Journal of Evolutionary Economics. Vol. 26 Issue 1.

Leal, S, Napoletano, M (2017), ‘Market Stability vs. Market Resilience: Regulatory Policies Experiments in an Agent-Based Model with Low- and High-Frequency Trading’, Journal of Economic Behavior and Organization.

Massad, T, (2015), Remarks of Chairman Timothy Massad before the Conference on the Evolving Structure of the U.S. Treasury Market. Available at: https://www.cftc.gov/PressRoom/SpeechesTestimony/opamassad-30

Menkveld, A (2013), ‘High Frequency Trading and the New-Market Makers’, Journal of Financial Markets, Vol. 16.

Moore, M, Schrimpf, A and Sushko, V (2016), ‘Downsize FX markets: causes and implications’, BIS Quarterly Review.

Noss, J, Pedace, L, Tobek, O, Linton, O and Crowley-Reidy, L (2017), ‘The October 2016 sterling flash episode’, Bank of England Staff Working Paper No. 687.

Salmon, C (2017), ‘Keeping up with fast markets’, speech at the 13th Annual Central Bank Conference on the Microstructure of Financial Markets.

Turrell, A (2016), ‘Agent-based models: understanding the economy from the bottom up’, Bank of England Quarterly Bulletin, p173-188.

SEC/CFTC (2010), ‘The market events of May 16, 2010: Report of the staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues’.

Zhao, H (2010), ‘Dynamic relationship between exchange rate and stock price: Evidence from China’, Research in International Business and Finance, Vol. 24, Issue 2.

Annex 1

Table A1 – Parameters specified for calibration exercise (those calibrated are in bold)

Parameter	Variable	Value
Model set-up		
T	Number of time periods	27,000
τ	Tick size (number of decimal places you are able to submit prices at)	5
$\boldsymbol{\pi_{MM_s}}$	**Probability of slow market maker being able to play at *t***	**[0.2,0.4]**
$\boldsymbol{\pi_{MM_f}}$	**Probability of fast market maker being able to play at *t***	**[0.2,0.4]**
π_{noise}	Probability of noise trader being able to play at *t*	0.02
$\boldsymbol{\pi_{fun}}$	**Probability of fundamental trader being able to play at *t***	**[0.1,0.2]**
$\boldsymbol{\pi_{mom}}$	**Probability of momentum trader being able to play at *t***	**[0,0.2]**
Market Makers behaviour		
ω_{MM_s}	Fraction of capital allowed at risk (slow)	0.2
ω_{MM_f}	Fraction of capital allowed at risk (fast)	0.12
γ	Market maker price sensitivity to realised volatility	3
$K_{MM_{s,0}}$	Initial capital of slow market maker	15000
$K_{MM_{f,0}}$	Initial capital of fast market maker	10000
$\bar{I}_{mm}$	Inventory limit	$0.5K_{MM_{k,0}}$
$\boldsymbol{\alpha}$	**Decay strength of Market maker price signal**	**[0.15,0.25]**
r	Reduction for adverse selection parameter	1000
n	Number of orders they post each side of the order book	4
$\underline{s}$	Minimum bound on inferring demand signal	0.00001
$\underline{\Delta}$	Minimum bound on bid-ask spread	0.0001
l_k	Window over which to consider realised price change	10
Noise trader behaviour		
σ_{noise}	Standard deviation of noise trades	50
Momentum trader behaviour		
$\boldsymbol{\omega_{mom}}$	**Momentum Order sensitivity parameter to realised price change**	**[50000 , 150000]**
h	Shape parameter	4
$\bar{I}_{mom}$	Inventory limit of momentum trader	300
l_{mom}	Window over which price change is considered	10
Fundamental trader behaviour		
σ_v	Standard deviation of fundamental price shock	0.0000156
ω_{fun}	Order sensitivity to deviation from fundamental price	500

Where possible, values have been taken directly from empirical data (such as the standard deviation of the price shock). Where this is not possible values have been assigned based on intuition