2

High-Frequency Trading in FX Markets

Anton Golub, Alexandre Dupuis, Richard B. Olsen
Olsen Ltd

This chapter provides an overview of the landscape and the basic mechanics of the foreign exchange (FX) markets and their organised exchanges. We explain algorithmic trading in the foreign exchange and analyse trading frequencies of different types of market participants. We continue with an overview of the key insights of academic literature of the impact of high-frequency (HF) traders in the foreign exchange market and discuss actual market events where there have been short-term price disruptions. We focus on the behaviour of the high-frequency traders involved.

There is definite empirical evidence of the path dependency of the price trajectory; a black swan event may be triggered at any time due to microstructure effects that are not linked to fundamental factors. Organised trading venues are exploring ways to prevent microstructure effects distorting price action, though without reaching a satisfactory solution so far. This chapter proposes a new method to achieve price stability. We suggest that the queuing system of limit order books rewards market participants by offering competitive two-way prices; model simulations presented here indicate that this might well enhance market stability.

THE CURRENCY MARKET

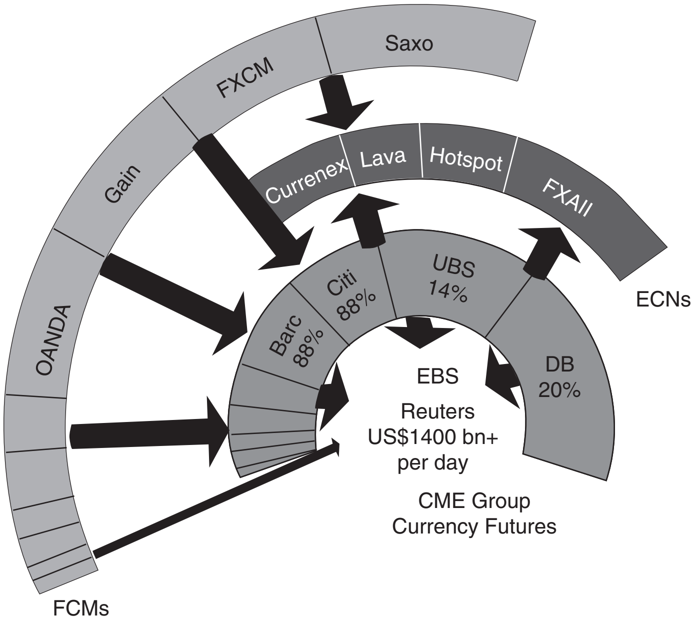
This section describes the currency market from a high-frequency trading (HFT) perspective. We give an overview of the overall landscape of the market and the relationships between the major players. The dynamics is illustrated by discussing their technical details. Then we review the trading algorithms that are used in the industry and, based on Schmidt (2011) and Masry (2013), we assess their impact on market microstructure.

Market venues

The currency market is a complex system of organised exchanges. At the centre of the market there are two inter-dealer electronic broking platforms: Electronic Broking Services (EBS) and Reuters. These platforms, described in some detail below, act as a source of interbank liquidity in the FX market and they are the place where large HFT players trade. The requirement of a minimum ticket size of one million units has created a business opportunity to build alternative trading venues for retail and other market participants. Currenex, for example, has built a multi-bank electronic communication network (ECN) and there is a similar platform available by Hotspot. These new ECNs provide their customers with more sophisticated tools for market making, such as full anonymity (Bank for International Settlements 2011), where the counterparty making a trade does not know who is taking the other side; this is not the case with EBS and Reuters. Retail aggregators have reshaped the market by providing small-scale traders with access to the foreign exchange markets. The retail aggregators have become a significant force in the overall foreign exchange market, making up approximately 10% of spot volume. Finally, the largest currency futures market is operated by Chicago Mercantile Exchange Group, with a daily volume of US$100 billion.

Figure 2.1 depicts how the different venues interact. We observe that the structure is not hierarchical, as Futures Commission Merchant (FCM) firms trade with large banks and possibly on EBS and Reuters. It is also interesting to note that Figure 2.1 is dynamic, and players may change their behaviour and reposition themselves within the foreign exchange market. An example is EBS, which has decided to decrease the minimal ticket size to 100,000 units for selected major currency pairs in order to attract smaller-scale traders.

Figure 2.1 Structure of the venues in the currency market



We now give some detail about some of the main venues from Figure 2.1.

EBS

EBS is the main venue for all the USD, EUR, GBP, CHF and JPY crosses. EBS provides two data feeds, one with time-sliced snapshots of the order book every 250 milliseconds[1] and a premium feed, EBS Live, which sends snapshots every 100 milliseconds. The snapshots were concurrently changed from showing the top level and two lower aggregated levels to showing ten levels of the limit order book. The minimum tick size was one pip, but was reduced to one tenth of a pip. That experiment proved to be unsuccessful; as of November 2012 the minimum tick size for most pairs reverted to one pip or half a pip. As already mentioned, EBS has a minimum ticket size of one million units and attracts large institutional traders. At the time of writing, they do not allow traders to modify orders or to have the last-look provision.[2] All quotes are pre-screened for credit, meaning that quotes will only be received from a given counterparty if the prime broker who actually clears the trade has the required credit line with the other counterparty or their prime broker. Most of the main pairs have a minimum quote lifetime (MQL) of 250 milliseconds, meaning that an order cannot be cancelled until 250 milliseconds have elapsed from the time it was added to the book. Ticks are not sent out on the data feed in a real-time manner, but are instead time-sliced to show an aggregate of the traded size at the best traded price over the interval. Filled quotes are reported immediately to the involved parties, before the rest of the market. Finally, we note that EBS has a multi-matching-engine architecture located in New York, London and Tokyo, and all engines operate independently but update one another when the order book is modified.

Reuters

Reuters is the main venue for all the crosses for Commonwealth currencies and Scandinavian currencies. Their rules are very similar to those at EBS summarised above. Reuters does not have multiple engines, and operates a single engine in London.

Currenex and Hotspot

Currenex has a fast architecture, which allows for streaming of order-based feeds and timely confirmations and executions. It does not provide an MQL feature, and no minimum ticket size is imposed. At Currenex, traders have the ability to modify orders instead of cancelling and replacing them, and to use conditional orders, execution algorithms and pegged orders. The tick size is 0.1 pips. Currenex does have some liquidity providers who use last-look provision. Hotspot is similar to Currenex, except that the minimum ticket size is 50,000 units and ticks are delayed by one second.

Hotspot does have liquidity providers who use last-look provision, though quotes from these traders can be filtered out. However, relative to the Currenex non-last-look feed, Hotspot's is relatively wide, suggesting the feature of allowing market participants to have a "last look" undermines liquidity and leads to wider spreads.

Oanda

Oanda is one of the major FCMs and one of the original FX dealers on the Internet.[3] The company's focus has been to build a highly scalable platform that executes transactions at minimum cost. In addition to the major currencies, Oanda offers trading in exotic exchange rates, precious metals and contracts for difference[4] of stock indexes and

US Treasuries. Transaction prices are identical for tickets as small as US$1 and as large as US$10 million, and the same across different market segments. Interest is paid on a second-by-second basis. Unlike traditional trading venues, Oanda offers firm quotes to its clients and hedges excess exposure with institutional market makers. Oanda's main revenue is generated from market making; it earns the spread between bid and ask prices at which its customers trade.

CME

The largest currency futures market is operated by the Chicago Mercantile Exchange (CME) Group,[5] with an average daily notional volume of approximately US$100 billion, most of it is being traded electronically. Similar to other futures products, currency futures are traded in terms of contract months with standard maturity dates typically falling on the third Wednesdays of March, June, September and December. The CME Group offers 49 currency futures contracts; the crosses of the G10 countries (ie, AUD, CAD, CHF, EUR, GBP, JPY, NOK, NZD, SEK, USD) as well as crosses of emerging markets, such as BRL, KRW and RMB. The minimum tick size is one pip, and the ticket size is US$125,000. Since 2010, CME Group has offered trading in selected E-micro FX futures, which are one-tenth of the standard size. In most cases, traders will offset their original positions before the last day of trading. Less frequently, contracts are held until the maturity date, at which time the contract is cash-settled or physically delivered, depending on the specific contract and exchange. Only a small percentage of currency futures contracts are settled in the physical delivery of foreign exchange between a buyer and a seller. The CME is responsible for establishing banking facilities in each country represented by its currency futures contracts, and these agent banks act on behalf of the CME and maintain a foreign currency account to accommodate any physical deliveries.

Unlike the FX spot market, CME provides a centralised pricing and clearing service, ie, the market price and the order book information for a currency futures contract will be the same regardless of which broker is used, and the CME guarantees each transaction. CME Group ensures that self-regulatory duties are fulfilled through its Market Regulation Department, including market integrity protection by maintaining fair, efficient, competitive and transparent markets.

Trading algorithms

We distinguish between two classes: algorithmic execution and algorithmic decision-making. The first addresses the automated execution of large orders in small tickets with the objective of minimising the price impact and/or ensuring the anonymity of execution. The second class groups the automated algorithms designed to generate Alpha.

When do we classify a trading algorithm as belonging to the class of high-frequency traders? Is the decisive criterion the number of trades? Or are there additional criteria? To shed some light on these questions, we follow Gomber *et al* (2011), who consider HFT as a subset of algorithmic trading (AT). HFT and AT share common features: pre-designed trading decisions, used by professional traders; observing market data in real-time; automated order submission; automated order management, without human intervention; use of direct market access. Gomber *et al* suggest criteria that only AT fulfil: agent trading; minimising market impact (for large orders); achievement of a particular benchmark; holding periods of possibly days, weeks or months; working an order through time and across markets.

Finally, we list criteria that only HFT satisfy: very high number of orders; rapid order cancellation; proprietary trading; profit from buying and selling; no significant positions at the end of day; very short holding periods; very low margins extracted per trade; low latency requirement; use of co-location/proximity services and individual data feeds; a focus on highly liquid instruments.

Indeed, as pointed out by Gomber *et al* (2011), HFT is not a trading strategy as such.

Algorithmic execution

The main idea of these strategies is to minimise the impact on price movement of buying or selling a large order. The algorithms at hand basically slice up the order into smaller orders and select appropriate times to transact in the hope that the average price will be close to the current price, that the impact will be low and that the trades will go unnoticed.

The basic algorithm is called time-weighted average price and slices time in an equal manner given a time horizon. This algorithm is not clever in the sense that it does not follow market activity and is easily detectable. An alternative is to define time as buckets of volume and allow the algorithm to trade a quantity when a given volume has been transacted in the market; in this case we talk about volume-weighted average price. Various improvements have been proposed to make these algorithms more adaptive to market activity and to include the effect of news. More details, as well as a classification, can be found in Almgren (2009) and Johnson (2010).

Algorithmic decision-making

We list the most common trading strategies designed to generate profit; in general they are believed to contribute to market liquidity (Chaboud *et al* 2012). The complexity of the algorithms and the large number of decisions that need to be taken implies that these algorithms need to be computerised and cannot be generated by hand. The algorithms are not discretionary decisions, as all of the responses of the trading model are predetermined. Depending on their implementations, these strategies can be classified either as HFT or not.

Market-making strategies are designed to offer temporary liquidity to the market by posting bid and ask prices with the expectation of earning the bid and ask spread to compensate for losses from adverse price moves. In some trading venues, these types of strategies are incentivised with rebate schemes or reduced transactions fees. We will explain later (see pp. 36ff) how such incentives can be used to make price discovery more robust and contribute to price stability.

Statistical arbitrage strategies are a class of strategies that take advantage of deviations from statistically significant market relationships. These relationships can, for example, be market patterns that have been observed to occur with some reasonable likelihood.

Mean reversion strategies assume that the price movement does not persist in one direction and will eventually revert and bounce back. This hypothesis is derived from the fact that positions eventually need to be closed, triggering a price reversal. An example of an automated strategy is described in Dupuis and Olsen (2012).

There exist arbitrage strategies to take advantage of price differences across platforms or take advantage of information ahead of delays. Traders embark, for example, on triangular arbitrage (eg, if buying x units of EUR/USD and selling x units of EUR/GBP, and

Table 2.1 Approximation of the percentage of filled trades per trading frequencies expressed in trades per day on EBS (Schmidt 2011)

Type	Frequency	Percentage
MT	—	1.9
Slow AI	<500	1.5
HFT	500–3000	2.3
Ultra-HFT	>3000	4.2

MT, manual trader; AI, automated interface.

selling the appropriate units of GBP/USD leads to an instantaneous and risk-free profit).

Liquidity detection strategies are used to spot large orders in the market and/or to trigger a particular behaviour by other market participants. These kinds of strategy are at the borderline of what is deemed ethical; examples are pinging (small orders to possibly hit hidden orders), quote stuffing (entering and immediately cancelling a large amount of orders to blur out the real state of the limit order book) or momentum ignition, where orders are placed to exacerbate a trend. These strategies are equivalent to spamming in the Internet; there is a need for subtle mechanism to minimise this type of abuse. We discuss such a mechanism below (see pp. 36ff).

Trading frequencies

To draw the landscape of the trading frequencies, we report on the results of Schmidt (2011) and Masry *et al* (2012), which have analysed transaction data from the currency market and provide the profile of the high-frequency traders with the durations of their positions.

Schmidt (2011) investigates transaction data from EBS, where, we recall, trading takes place through a limit order book and where the minimal order size is of US$1 million, which makes it a trading venue for large players. The study by Masry *et al* (2012) analyses transaction data from the market maker Oanda that, in contrast to EBS, has largely retail and some institutional investors; all traders have the same price terms for transactions from US$1 up to US$10 million.

Schmidt (2011) describes in detail the composition of the traders at EBS based on transaction data during the six months between May 2011 and the end of November 2011. The study defines two types

Table 2.2 Percentages of trades that were executed at various frequencies f on the Oanda platform, as computed by Masry *et al* (2013)

Frequency	Percentage
$f \leqslant 50$	27.7
$50 < f \leqslant 100$	7.9
$100 < f \leqslant 500$	12.5
$f > 500$	52.0

Frequencies f are expressed in trades per day. Transaction data spans January 2007–March 2009.

of traders: manual traders (MTs), who use a graphic-user-interface-based (GUI-based) based access, and automated traders, who use an automated interface (AI) for trading. AI trading is further subdivided into three subcategories: "Slow AI" at less than 500 trades per day, "HFT" at between 500 and 3000 trades per day and "Ultra-HFT", with more than 3000 trades per day. MTs account for 75% of EBS customers and more than 90% of MTs submit on average less than 100 orders a day. For EUR/USD, the currency pair with the largest volume, the average number of daily orders submitted by EBS customers are: MT 3.7%; Slow AI 5.7%; HFT 29%; Ultra-HFT 61.6%. Other currency pairs show similar patterns. The difference between Ultra-HFT and MT appears to be massive at first sight, but this neglects the fact that high-frequency traders typically cancel a large percentage of their orders. Schmidt (2011) reports that the average fill ratio for various EBS customer groups is around 50% for MTs, and (considering only AI belonging to the professional trading community; see Schmidt 2011) 26.6% for Slow AI, 8.1% for HFT and 6.8% for Ultra-HFT. Using the above numbers, we approximate the percentage of filled trades per trading frequency and show these in Table 2.1.

Masry *et al* (2013) analyse the transactions done at Oanda between January 2007 and March 2009. The data set comprises 110 million transactions belonging to 46,000 different accounts acting in 48 different currency pairs. First Masry *et al* categorised the traders, assigning them a trading frequency by computing the average number of transactions they made per day. Note that special care was taken to differentiate between accounts with different trading frequencies.

Percentage shares of the different trading frequencies are shown in Table 2.2.

Assuming that manual traders can trade at most 50 times a day, the corresponding percentage matches the one on EBS (26.6%). This proportion is suggested by Tables 2.1 and 2.2.

ACADEMIC LITERATURE

With the rapid increase of trading volume from HFT, academic studies have investigated how computerised trading affects the overall market quality. Cvitanic and Kirilenko (2011) derive theoretical distributions of transaction prices in limit order markets populated by low-frequency traders (humans) before and after the entrance of a high-frequency trader (machine). They find that the presence of a machine is likely to change the average transaction price and that the distribution of transaction prices has more mass around the centre and thinner tails. Jarrow and Protter (2011) express concern that the speed advantage of HF traders and the potential commonality of trading actions among computers may have a negative effect on the informativeness of prices. This is because computerised traders, triggered by a common signal, collectively act as one big trader, giving rise to price momentum, causing prices to be less informationally efficient. Cespa and Foucault (2012) argue that the self-reinforcing relationship between price informativeness and liquidity is a source of contagion and fragility: a small drop in the liquidity of one security propagates to other securities and can, through a feedback loop, result in a large drop in market liquidity. This leads to multiple equilibria characterised by either high illiquidity and low price informativeness or low illiquidity and high price informativeness, where the former type of equilibrium generates a liquidity crash similar to the Flash Crash on May 6, 2010.

Empirical academic research has mostly focused on the effects of HFT on the market quality in equity markets. The studies have shown that, in general, computerised trading improves traditional measures of market quality and contributes to price discovery. Hendershott *et al* (2011) study the 30 largest DAX stocks on the Deutche Boerse and find that AT represents a large fraction of the order flow and contributes more to price discovery than human traders. Algorithmic traders are more likely to be at the inside quote when spreads are high than when spreads are low, suggesting that algorithmic traders supply liquidity when this is expensive and demand liquidity when this is cheap. Hendershott *et al* find no evidence that AT increases volatility. Hendershott and Riordan (2011) examine the impact AT has on the market quality of NYSE listed stocks. Using a normalised measure of NYSE message traffic surrounding the NYSE's implementation of automatic quote dissemination in 2003, they find AT narrows spreads, reduces adverse selection and increases the informativeness of quotes, especially for larger stocks. Hasbrouck and Saar (2012) measure HFT activity by identifying "strategic runs" of submission, cancellations and executions in the Nasdaq order book. They find that HFT improves market quality by reducing short-term volatility, spreads and depth of the order book. Menkveld (2012) claims a large high-frequency trader provides liquidity, and its entrance into the market leads to a decrease in spreads. Brogaard (2010) examines the impact of HFT on the US equity market using a unique HFT data set for 120 stocks listed on Nasdaq. HFT is found to add to price discovery, providing the best bid and offer quotes for a significant portion of the trading day, and reducing volatility. However, the extent to which HFT improves liquidity is mixed, as the depth high-frequency traders provide to the order book is a quarter of that provided by non-high-frequency traders.

The most detailed examination of the impact of HFT on the FX market was made by Chaboud *et al* (2012) using high-frequency trading data from EBS for the period September 2003–September 2007 in three exchange rates: EUR/USD, USD/JPY and EUR/JPY. The crucial feature of their data set is that, on a minute-by-minute frequency, the volume and direction of human and computer trades are explicitly identified, allowing explicit measurement of the impact of high-frequency traders. They find very strong evidence that computers do not trade with each other as much as predicted, concluding that the strategies used by algorithmic traders are more correlated and less diverse than those used by human traders.

Next, they investigate the effect that both algorithmic trading activity and the correlation between algorithmic trading strategies have on the occurrence of triangular arbitrage opportunities. They indicate that algorithmic trading activity is found to reduce the number of triangular arbitrage opportunities, as the algorithmic traders quickly respond to the posted quotes by non-algorithmic traders and profit from any potential arbitrage.

Furthermore, a higher degree of correlation between algorithmic trading strategies reduces the number of arbitrage opportunities. There is evidence that an increase in trading activity where computers are posting quotes decreases the number of triangular arbitrage opportunities. Algorithmic traders make prices more efficient by posting quotes that reflect new information.

Chaboud *et al* also investigate the effect algorithmic traders on the degree of autocorrelation in high-frequency currency returns: they estimate the autocorrelation of high-frequency, five-second returns over five-minute intervals. Similar to the evolution of arbitrage opportunities in the market, the introduction and growth of algorithmic trading coincides with a reduction in the absolute value of autocorrelation. On average, algorithmic trading participation reduces the degree of autocorrelation in high-frequency currency returns by posting quotes that reflect new information more quickly.

Finally, Chaboud *et al* report highly correlated algorithmic trading behaviour in response to an increase in absolute value of the autocorrelation in high-frequency currency returns; this supports the concern that high-frequency traders have very similar strategies, which may hinder the price discovery process (Jarrow and Protter 2011).

HFT during time of market stress

The availability of liquidity has been examined in equity markets; academic studies indicate that, on average, high-frequency traders provide liquidity and contribute to price discovery. These studies show that high-frequency traders increase the overall market quality, but they fail to zoom in on extreme events, where their impact may be very different. A notable exception is the study by Kirilenko *et al* (2011) that uses audit-trail data and examines trades in the E-mini S&P 500 stock index futures market during the May 6, 2010, Flash Crash. They conclude that high-frequency traders did not trigger the Flash Crash; HFT behaviour caused a “hot potato” effect and thus exacerbated market volatility.

In contrast to these studies, the following sections provide anecdotal evidence of the behaviour of computerised traders in times of severe stress in foreign exchange markets:

- the JPY carry trade collapse in August 2007;

- the May 6, 2010, Flash Crash;
- JPY appreciation following the Fukushima disaster;
- the Bank of Japan intervention in August 2011 and Swiss National Bank intervention in September 2011.

While each of these episodes is unique in terms of the specific details and they occurred at different stages of the evolution of high-frequency traders, these events provide valuable insight into how computerised traders behave in periods of large price moves.

August 2007 yen appreciation

The August 16, 2007, USD/JPY price rise was the result of the unwinding large yen carry-trade positions; many hedge funds and banks with proprietary trading desks had large positions at risk and decided to buy back yen to pay back low-interest loans. Chaboud *et al* (2012) provide details of this event, and report that the event had one of the highest realised volatilities and the highest absolute value of serial correlation in five-second returns. The yen appreciated sharply against the US dollar at around 06h00 and 12h00 (New York time). The two sharp exchange rate movements happened when trading algorithms, as a group, aggressively sold dollars and purchased yen; at the other side of these trades were human traders, not other algorithms. Human traders were selling and buying dollars in almost equal amounts. The orders initiated by computers were more correlated than the than those of humans. After 12h00, human traders, in aggregate, began to buy dollars fairly aggressively, and the appreciation of the yen against the dollar was partly reversed.

Flash Crash, May 6, 2010

On May 6, 2010, the US stock market experienced one of its biggest price drops, with the Dow Jones Industrial Average (DJIA) index losing 900 points in a matter of minutes. It was the second largest intraday point swing, 1010.14 points, and the biggest one-day point decline, of 998.5 points. Such a large swing raised concerns about the stability of capital markets, resulting in a US Securities and Exchange Commission (SEC) investigation (US Securities and Exchange Commission and the Commodity Futures Trading Commission 2010). This report claimed that the crash was triggered by a sell algorithm of a large mutual fund executing a US$4.1 billion sell trade in the

E-mini S&P 500 futures, and while HFT did not spark the crash, it does appear to have created a "hot potato" effect contributing to the crash. Nanex (2010) reported that quote saturation and NYSE Consolidated Quotation System (CQS) delays, combined with negative news from Greece together with the sale of E-mini S&P 500 futures "was the beginning of the freak sell-off which became known as the Flash Crash". Menkveld and Yueshen (2013) analysed the May 6, 2010, Flash Crash using public and proprietary trade data on E-mini S&P 500 futures and S&P 500 Exchange Traded Fund (ETF) and found that the large mutual fund, whose E-mini trading reportedly contributed to the crash, was relatively inactive during the period of the crash, as its net selling volume was only 4% of the total E-mini net sells.

Sharp price movement was also witnessed in the FX market. Analysing the data from EBS, Bank for International Settlements (2011) showed that algorithmic execution comprised about 53.5% of total activity, versus 46.5% manual, which was higher than on average (45% algorithmic, 55% manual for 2010), suggesting that algorithmic participants did not reduce activity, as was the case for traditional market participants. The price movement is compared against two additional measurements, the ratio of algorithmic investor order submissions on May 6 to average algorithmic investor order submissions for the prior period, and the ratio of manual investor order submissions on May 6 to average manual investor order submissions for the prior period. Both manually and algorithmically submitted orders were in fact much higher than the average of the prior period. The share of algorithmic activity generated by the professional trading community (PTC) as a share of total algorithmic activity was higher than the average, suggesting that the increased contribution of algorithmic participants was driven largely by the increased activity of PTC participants.

March 2011 yen appreciation

Early in the morning of March 17, 2011, in the days following the Fukushima Daiichi earthquake, the USD/JPY declined by 300 pips, from around 79.50 to below 76.50 in just 25 minutes, between 05h55 and 06h20 Tokyo time (16h55–17h20 New York time on March 16, 2011). This price movement was triggered by stop-loss trades of retail FX margin traders (Bank for International Settlements 2011). The margin calls that the retail aggregators executed on behalf of their traders set off a wave of USD selling in a thin market. Many banks withdrew from market making and others widened their spreads so much that their bids were far below the last prevailing market price. This created a positive feedback loop of USD/JPY falling and leading to even more stop-losses until the pair hit 76.25 at around 06h20. The exchange rate recovered in the next 30 minutes to 78.23 as hedge funds and new retail investors began to build up fresh long positions. Banks, having withdrawn from making prices during the most volatile period, resumed market making. The USD/JPY dropped again at around 07h00, to reach 77.10, coinciding with another round of automated stop-outs, executed this time by the FX margin-trading brokers that participate on a particular trading platform on the Tokyo Futures Exchange. When the system restarted at 06h55, numerous compulsory stop-out orders were generated over five minutes to the six market makers that have obligations to provide prices to this platform (an estimated US$2 billion of USD/JPY selling). During this episode, both high-frequency traders and traditional market makers withdrew from the market.

The episode suggests that, even in trading venues with designated market makers, there is no guarantee of the quality of the quotes, as some market makers with formal obligations to quote prices widened their bid–offer spread considerably during that time.

Central Bank interventions

This section discusses the behaviour of high-frequency traders during central banks interventions. We focus on two events: the Bank of Japan (BOJ) intervention on August 4, 2011, following the Fukushima Daiichi earthquake, and the Swiss National Bank (SNB) intervention on September 6, 2011, following a strong appreciation of Swiss franc. As mentioned previously (see page 29), Schmidt (2011) separates traders into manual traders, who use EBS's proprietary GUI access for order management, and automated traders, who use AI for trading (slow AI users, high-frequency traders and ultra-high-frequency (UHF) traders). Two liquidity measures are calculated: the percentage of time that traders of each group provide two-sided liquidity, and bid–offer spread compiled on a one-second grid and averaged over 10-minute time intervals, with USD/JPY, EUR/JPY currency pairs for BOJ intervention, and EUR/CHF, USD/CHF currency pairs for SNB intervention.

The BOJ intervention at 01h00 GMT on August 4, 2011, caused a sharp jump in the USD/JPY exchange rate that did not disrupt the two-sided market. MTs provided liquidity 100% of the entire intervention time, while HF traders and slow AI failed to provide two-sided liquidity only for two seconds and eight seconds, respectively. UHF traders provided only intermittent liquidity during first the 10 minutes after intervention and withdrew from the market for several minutes around 02h40 GMT. The spread was always determined by HFT, while the spread formed by slow AI users after the intervention was wider than that of MT users, implying slow AI may be even more risk averse than MTs.

The SNB intervention on September 6, 2011, lasted for 30 minutes, from 08h00 to 08h30 GMT. Liquidity provided by slow AI users for the USD/CHF exchange rate had notable gaps prior to the SNB intervention. The intervention briefly decreased the percentage of time for which all customer groups quoted two-way prices. HF traders were the quickest, while UHF traders were the slowest in restoring the two-sided market. HF traders were the most active in setting the bid–offer spread during and after the intervention. For the EUR/CHF exchange rate, MTs and HF traders were the best liquidity providers during the SNB intervention. While the SNB intervention affected the EUR/USD exchange rate, its liquidity was not impaired.

These two events suggest that high-frequency traders can be valuable contributors to market liquidity during dramatic price moves of exchange rates, such as during central bank interventions. In other scenarios, HF traders can also destabilise price action, because they may be forced to close out positions all of a sudden, thus triggering an avalanche. In the next section, we shall discuss how organised exchanges can improve price discovery and reduce the likelihood of a "flash crash".

ALTERNATIVE LIMIT ORDER BOOK

Price action in financial markets is at times erratic, because second-by-second transaction volume is a mere trickle, and minor market orders can trigger a price spike that can set off a large price move due to margin calls. Price movements are spurious and respond in a nonlinear fashion to imbalances of demand and supply. A temporary reduction in liquidity can easily result in significant price moves, triggering stop losses and cascades of position liquidations. High-frequency traders now account for a large share of total transaction volume; if these traders are taken by surprise and close out their positions in one go, then this can trigger a massive sell-off, akin to the May 6, 2010, Flash Crash.

In response to the Flash Crash and other similar events, regulators have introduced several rules to ensure orderly functioning of capital markets. Market-wide circuit breakers, the so-called "Limit Up–Limit Down", have been put in place in US equity markets, to halt trading in the case of violent price moves (US Securities and Exchange Commission 2012). European regulators went a step further and burdened trading firms that use HFT with several new trading obligations (European Commission 2011). First, high-frequency traders are required to provide two-sided liquidity on a continuous basis, regardless of the prevailing market conditions. Second, all orders submitted by high-frequency traders will be obligated to stay in the order book for at least 500 milliseconds. Orders placed in the order book cannot be cancelled or changed during that predefined time frame. Exchanges cap the number of orders that high-frequency traders can submit or charge them additional costs. The most extreme form of regulation is the so-called Tobin Tax, a small fee on transactions of financial securities. France is the first European country to impose such a transaction tax, which amounts to 0.2%, to be paid on all transactions by companies headquartered in France. Becchetti *et al* (2013) analysed the impact of the introduction of the French Tobin tax on volume, liquidity and volatility of affected stocks and documented that the tax has a significant impact in terms of reduction in transaction volumes and intraday volatility.

High-frequency traders are required to invest in superior technology and sophisticated trading models, risking their own capital, while providing ample liquidity and performing valuable service to market participants. Regulators and operators of organised exchanges have imposed additional costly obligations on high-frequency traders; there has been little discussion on what incentives are necessary to induce liquidity providers to stay in the market during stressful periods. We believe that the limit order queuing mechanism needs to reward competitive two-sided limit orders and give them preferential queuing status over one-sided limit orders.

Therefore, in the rest of this section we propose an order book mechanism that combines price ranking with spread ranking to queue limit orders, which we call spread/price–time priority. We use the agent-based model by Bartolozzi (2010) to analyse the benefits of the aforementioned priority mechanism. The simulations provide evidence that the spread/price–time priority is successful in increasing the overall market quality.

Spread/price–time priority

Most modern trading venues operate under a price–time priority mechanism. Price-time priority determines how limit orders are prioritised for execution. The primary priority is price: the lowest sell limit order (offer) is the first to receive execution against market buy orders, while the highest buy limit order (bid) is the first to receive execution against market sell order. The secondary ranking attribute is the time at which a limit order has been submitted to the order book. We propose an order queuing mechanism based on spread/price–time priority, where the ranking mixes the price ranking and the spread ranking according to a parameter $\alpha \in [0, 1]$.

Liquidity providers submit limit orders and those limit orders can be either one sided (ie, a submitted limit order is either a buy or sell limit order) or two sided, in which case the trader simultaneously submits a buy and a sell limit order. The limit order is assigned with a rank, $\text{rank}(\alpha)$.

We propose an alternative set-up. We want to reward market participants, who reveal information not only about the trade that they want to do, but also about the other side of the trade. If a trader wants to sell, we do not rank their order only on the basis of their sale price, but also on the size of the spread: how far away the trader sets the ask. This is valuable information for price discovery; if the spread is narrow, then the market maker has a balanced expectation; if the spread is wide, then their expectation is skewed.

The queuing of limit orders within the order book is done according to a weighted average between a price contribution (weight α) and a spread contribution (weight $1 - \alpha$). In other words, the rank of an limit order equals

$$\text{rank}(\alpha) = \alpha \times \text{price} + (1 - \alpha) \times \text{spread}$$

The buy/sell limit order with lowest rank receives the highest priority for execution against sell/buy market orders. The price rank of limit orders is computed as the price difference from the currently resting limit order with best price. If the price of the newly submitted limit order sets the new best price, then the price rank of limit order will equal zero. Limit orders resting on the same side will update their price rank according to the newly set best price.[6] In other words, price ranking of a submitted limit order equals

$$\text{price} = \begin{cases} \text{ask} - \text{ask}^{\text{best}}, & \text{sell limit order} \\ \text{bid}^{\text{best}} - \text{bid}, & \text{buy limit order} \\ 0, & \text{new best buy/sell price} \end{cases}$$

where ask^{best} and bid^{best} are the best selling price and best buying price of resting limit orders. The spread ranking of two-sided limit order is computed as the difference between the price of buy limit order and sell limit order. One-sided limit orders have the same spread ranking as the resting two-sided limit order with worst spread. In other words

$$\text{spread} = \begin{cases} \text{ask} - \text{bid}, & \text{two-sided limit order} \\ \text{spread}^{\text{max}}, & \text{one-sided limit order} \end{cases}$$

where $\text{spread}^{\text{max}}$ is the largest spread of currently resting two-sided limit order. Therefore, the spread ranking of one-sided limit orders are "at par" as spread ranking of resting two-sided limit order with worst spread. Finally, if the limit orders have the same rank, $\text{rank}(\alpha)$, time priority determines the queuing position. We note that the parameter α is used to "tune" the significance of spread versus price for primary ranking. For instance, decreasing the α parameter puts more weight to ranking based on spread, therefore providing a bigger incentive for traders to submit two-sided limit orders to the order book as these limit orders will have greater priority for execution. On the other hand, increasing the α parameter puts more weight to ranking based on price, therefore providing incentive for traders to submit price competitive limit orders. Note that setting $\alpha = 1$ reduces the limit order queuing mechanism to price–time priority, while setting $\alpha = 0$ reduces the limit order queuing mechanism to spread–time priority, where the price plays no role at all in queueing limit orders. Finally, we note that it is possible for buy and sell limit orders to "cross" or "lock" in price, for parameter α larger than zero.[7]

Agent-based model

We have used the agent-based model by Bartolozzi (2010) to evaluate the impact of spread/price–time priority ranking of limit orders on the market quality. The agent-based model has shown to be able to reproduce several empirical features of the high-frequency dynamics of the market microstructure: negative autocorrelation in returns, clustering of trading activity (volatility, traded volume and bid–ask spread), non-linear response of the price change to the traded volume, as well as average shape of the order book and volume imbalances. We shall briefly present the model; for the details we refer the reader to Bartolozzi (2010).

The market model evolves in discrete time steps, during which agents may undertake a certain action or just wait for a more profitable opportunity, ie, cancellation or active trading, the latter including both limit and market orders. All decision steps are based on dynamical probabilities, which are functions of private and public information. At each step, specifications for each order, such as type (limit or market), price (for limit orders) and volume are decided. The agents have access to the current state of the limit order book: all of the smoothed indicators are derived by this knowledge, such as the exponential midpoint price or volatility, and are classified as public information. Private information is represented by a simple Gaussian process, independent for each trader, with zero mean and standard deviation proportional to the volatility of the market. The limit order is automatically removed if it has not been executed in a certain number of time increments, or according to a strategic decision based on the current market condition, whereas it is more likely to cancel the limit order in a more volatile market. Agents with no orders in the order book evaluate the possibility of entering the market and their decision is based on a stochastic variable that represents the "level of confidence" in their price forecast, ie, market sentiment, which relates the public and the private information. The market sentiment can be thought of as the convolution between the agents, their trading strategies, the private information and the risk factors evaluated via the public information: the stronger the signal, the more likely it is that the trader takes a decision. If the agent enters the market, the type of the order is decided based on its relative position to the best prices: if the resulting submission price is greater than the ask price and the order is long (or lower than the bid price and the trade is short), then this is interpreted as a market order, while all the other orders are considered limit orders.

Our contribution to the agent-based model is the spread/price–time priority ranking of limit orders and the ability of traders to post two-sided limit orders, ie, to be market makers. If the trader decides to post a limit order, they will post a two-sided limit order with probability $p \in [0, 1]$, or they will a post one-sided limit order with probability $1 - p$. In the case when the trader decides to post a two-sided limit order, they will do so by maintaining a spread of at most 10 ticks (in real life, this amounts to a spread of at most 10 pips). Therefore, the agent-based model has two degrees of freedom that are left for user input: the parameter α, determining the primary ranking $\text{rank}(\alpha)$ in spread/price–time priority, and parameter p, determining the probability of submitting a two-sided limit order.

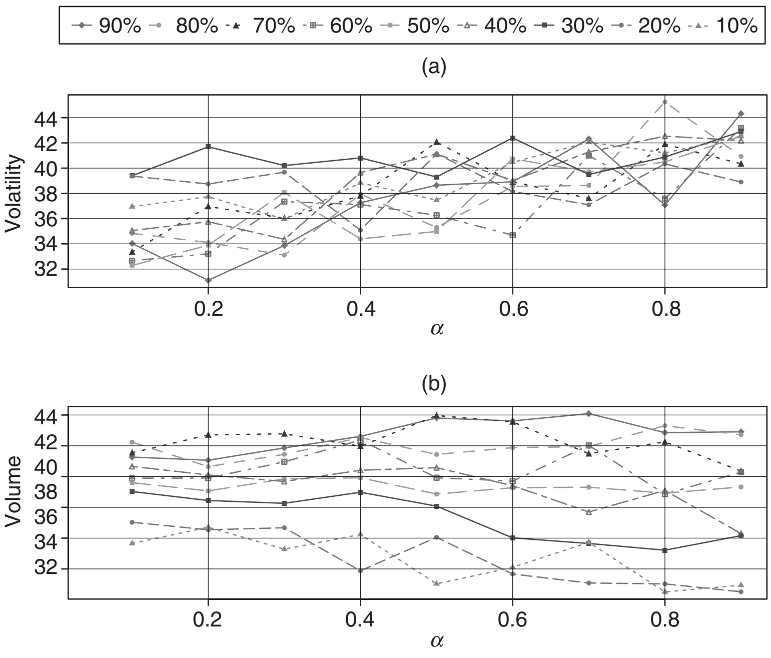
Results and conclusion

In this section we present the results of the agent-based model simulations, claiming that spread/price–time priority ranking is suitable for decreasing the market volatility, while not affecting the overall volume (this effect on volume is likely to occur with a brute-force mechanism such as a Tobin tax). Therefore, spread/price–time priority ranking provides benefits to both short- and long-term traders: high-frequency traders would keep their source of revenue from market making, while long-term investors would be able to operate in a stable market environment.

The agent-based model has two degrees of freedom: the parameter $\alpha \in [0, 1]$ determining the primary ranking $\text{rank}(\alpha)$, and the parameter $p \in [0, 1]$ determining the probability of submission of two-sided limit orders. Both of these parameters are chosen on a linear grid of nine values, ranging from 0.1 to 0.9. Each simulation has 1000 iterations and, running it for all pairs of parameters (α, p), we obtained a total of 81 simulation runs. Our primary concern was to analyse if the spread/price–time ranking was successful in decreasing the price volatility while not reducing the overall volume.

We compute price volatility as an average of price volatilities computed at smaller and larger scales. In other words, price volatility σ is an average of the price volatility for 1, 2, 5, 10, 15, 20, 25 and

Figure 2.2 (a) Average volatility and (b) average volume of agent-based model for a linear grid of probabilities (p = 10–90%) and parameter α



50 time steps, δt

$$\sigma = \frac{\sigma_{\delta t} + \sigma_{2\delta t} + \sigma_{5\delta t} + \sigma_{10\delta t} + \sigma_{15\delta t} + \sigma_{20\delta t} + \sigma_{25\delta t} + \sigma_{50\delta t}}{8} \quad (2.1)$$

where $\sigma_{n\delta t}$ is the standard deviation of returns computed after $n\delta t$ iterations. In this manner we have included timescales of interest to both high-frequency and low-frequency traders. The total volume is computed as the total turnover in a simulation run.

Figure 2.2 shows the average price volatility and total volume. Part (a) shows the price volatility, while part (b) shows the total volume, both as a function of α determining priority ranking rank(α) and as a function of p determining the probability of submitting a two-sided limit order. Figure 2.2 indicates that decreasing α, ie, putting more weight on the spread, results in lower volatility regardless of the probability p of submitting a two-sided order. Furthermore, it is clear that the total volume is highly dependent on parameter p determining the probability of submitting a two-sided limit order, where the higher probability will result in greater turnover of volume. On the other hand, there does not seem to be an obvious relationship between the resulting volatility and turnover of volume. In summary, the lower volatility obtained with spread/price–time does not necessarily lead to a loss of volume.

CONCLUSION

The foreign exchange market with a daily spot transaction volume of US$1.4 trillion is at the core of the global economy; the foreign exchange market sets the exchange rates between countries and is decisive for the health of the economic system. Today, there is significant evidence that the spuriousness of price action can cause a major price cascade in the foreign exchange markets, akin to the flapping of the wings of a butterfly causing a tornado. These price distortions do not wash out with the other economic uncertainties. A price crash in the foreign exchange market can destabilise the economic system even further. We review clear evidence that high-frequency trading provides liquidity in dramatic events, acknowledging that some trading practices are on the verge of being unethical. To minimise the occurrence of such practices, we have suggested an alternative queuing system for organised exchanges that rewards market makers for revealing private information and providing liquidity on an ongoing basis. The simulations indicate that price quality improves significantly, without the dramatic impact on volume that might well occur with the introduction of Tobin tax. There remain many open questions in the detailed mechanics of how price shocks are propagated. Researchers need access to comprehensive data sets so that they can study in detail the market mechanics and get a deeper understanding of the complex feedback processes. It is necessary to discuss and analyse alternative queuing systems in order to develop market mechanisms that are robust and ensure consistent pricing, independent of random variations of supply and demand. Financial markets are the equivalent of bridges in a transport system; they need to be stable and robust to rapidly changing buy and sell flows to be optimal for the economic system as a whole.

[1] One millisecond is one thousandth of a second.

[2] The last-look provision is a waiting period of several hundred milliseconds in which the liquidity provider has an option to fill or pass the incoming market order.

[3] See http://www.oanda.com.

4 Contract for difference is a financial derivative that allows traders to speculate on price movement of the underlying instrument, without the need for ownership of the instrument.

5 Smaller currency futures markets are present worldwide, including NYSE Euronext, the Tokyo Financial Exchange and the Brazilian Mercantile and Futures Exchange.

6 The updating process will manifest as adding a constant to ranks of all limit orders resting on the same side, ie, $\text{rank}(\alpha)_i = \text{rank}(\alpha)_i + \text{const}$, $i = 1, \dots, n$, and it will not change the queuing of resting limit orders regardless of the parameter $\alpha \in [0, 1]$.

7 Crossed quotes occur in a given security when the best buying price is higher than the best selling price. Locked quotes occur when the best buying price is equal to the best selling price.

REFERENCES

Almgren, R., 2009, "Quantitative Challenges in Algorithmic Execution", Presentation, http://www.finmath.rutgers.edu/seminars/presentations/Robert%20Almgren_37.pdf.

Bank for International Settlements, 2011, "High-Frequency Trading in the Foreign Exchange Market", Report, September.

Becchetti, L., M. Ferrari and U. Trenta, 2013, "The Impact of the French Tobin Tax", CEIS Research Paper 266, March.

Bartolozzi, M., 2010, "A Multi Agent Model for the Limit Order Book Dynamics", *The European Physical Journal B* 78(2), pp. 265–73.

Brogaard, J., 2010, "High Frequency Trading and Its Impact on Market Quality", Technical Report, July.

Cespa, G., and T. Foucault, 2012, "Illiquidity Contagion and Liquidity Crashes", Working Paper, May.

Chaboud, A., E. Hjalmarsson, C. Vega and B. Chiquoine, 2012, "Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market", Technical Report, October.

Cvitanic, J., and A. A. Kirilenko, 2011, "High Frequency Traders and Asset Prices", Technical Report, March.

Dupuis, A., and R. B. Olsen, 2012, "High Frequency Finance: Using Scaling Laws To Build Trading Models", in J. James, I. W. Marsh and L. Sarno (eds), *Handbook of Exchange Rates*. Chichester: Wiley Finance.

European Commission, 2011, Markets in Financial Instrument Directive 2. Technical Report, October.

Gomber, P., B. Arndt, M. Lutat and T. Uhle, 2011, "High-Frequency Trading", Technical Report Commissioned by Deutsche Boerse Groupe.

Hasbrouck, J., and G. Saar, 2012, *Low-Latency Trading*, Johnson School Research Paper Series no. 35-2010, December.

Hendershott, T., and R. Riordan, 2011, "Algorithmic Trading and Information", Technical Report, June.

Hendershott, T., C. M. Jones, and A. J. Menkveld, 2011, "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance* 66(1), pp. 1–33.

Jarrow, R., and P. Protter, 2011, "A Dysfunctional Role of High Frequency Trading in Electronic Markets", Technical Report, Cornell University Working Paper, June.

Johnson, B., 2010, *Algorithmic Trading and DMA: An Introduction to Direct Access Trading Strategies*. London: 4Myeloma Press.

Kirilenko, A. A., A. S. Kyle, M. Samadi and T. Tuzun, 2011, “The Flash Crash: The Impact of High Frequency Trading on an Electronic Market”, Technical Report, May.

Masry, S., 2013, “Event Based Microscopic Analysis of the FX Market”, PhD Thesis, University of Essex.

Menkveld, A., 2012, “High Frequency Trading and the New-Market Makers”, Technical Report, February.

Menkveld, A. J., and B. Z. Yueshen, 2013, “Anatomy of the Flash Crash”, SSRN Working Paper, April.

Nanex, 2010, “May 6th 2010 Flash Crash Analysis: Final Conclusion”, August, http://www.nanex.net/FlashCrashFinal/FlashCrashAnalysis_Theory.html.

Schmidt, A., 2011, “Ecology of the Modern Institutional Spot FX: The EBS Market in 2011”, Technical Report, Electronic Broking Services.

US Securities and Exchange Commission, 2012, “SEC Approves Proposals to Address Extraordinary Volatility in Individual Stocks and Broader Stock Market”, Press Release, June.

US Securities and Exchange Commission and the Commodity Futures Trading Commission, 2010, “Finding Regarding the Market Events of May 6”, Technical Report, September.