Government
Office for
Science

Foresight

What is the economic impact of the MiFID rules aimed at regulating high-frequency trading?

In association with

Oxera

An economic impact assessment

This paper has been commissioned as part of the UK Government's
Foresight Project on The Future of Computer Trading in Financial Markets.
The views expressed are not those of the UK Government and do not represent its policies.

Contents

1 Introduction and executive summary

1.1 Objectives and remit

Foresight has commissioned Oxera to conduct an economic impact assessment of the rules proposed by the European Commission in its review of MiFID aimed at regulating high-frequency trading (HFT).[1]

In recent years, the amount of trading done by 'high-frequency traders' has increased significantly, resulting in a debate about the impact on society of this type of algorithmic trading (AT). On the one hand, there may be benefits in terms of increased liquidity; on the other hand, it may facilitate the implementation of certain trading strategies that could be considered harmful or abusive. HFT may also have consequences for financial stability. Furthermore, it has been argued that the provision of additional liquidity may, in practice, be more limited than initially thought, and that high-frequency traders may take liquidity from, rather than provide it to, long-term investors, particularly at times when liquidity is already low and/or the market in a particular security is under stress.

As part of the MiFID review, the Commission has proposed a number of measures to address potential concerns about the impact of AT and HFT:[2]

1) A series of new specific organisational requirements for market participants would be introduced with the possibility of further specification in implementing acts on each of the issues below:

- authorised firms involved in automated trading would have in place robust risk controls to mitigate potential trading system errors;
- firms involved in automated trading would notify their competent authority of the computer algorithm(s) they employ, including an explanation of its design, purpose and functioning;
- firms who provide 'sponsored access' to automated traders would have in place robust risk controls and filters to detect errors or attempts to misuse facilities;
- operators of trading venues would have in place proper risk controls and arrangements to mitigate the risk of errors generated by automated trading leading to disorderly trading (e.g. circuit breakers) or the breakdown of their trading systems (e.g. by stress testing to ensure resilience);
- operators of trading venues would give equal and fair access to market participants to co-location services.

2) Implementing measures could further specify minimum tick sizes;

3) Market operators would be required to ensure that if a high frequency trader executes significant numbers of trades in financial instruments on the market then it would continue providing liquidity in that financial instrument on an ongoing basis subject to similar conditions that apply to market-makers; and

[1] European Commission (2010), 'Public Consultation: Review of the Markets in Financial Instruments Directive (MiFID)', Consultation Document, European Commission, December 8th.

[2] These rules would be applied to AT, which would be defined in a broad manner as 'trading involving the use of computer algorithms to determine any or all aspects of the execution of the trade such as the timing, quantity and price.' HFT would be considered a subcategory of AT. Furthermore, all persons involved in HFT over a specified minimum quantitative threshold would need to be authorised as investment firms. This would ensure that they are subject to organisational requirements (such as systems and risk management obligations and capital requirements) and to full regulatory oversight.

> 4) Market operators would be required to ensure that orders would rest on an order book for a minimum period before being cancelled. Alternatively they would be required to ensure that the ratio of orders to transactions executed by any given participant would not exceed a specified level. In either case, further specification would be needed on the specific period or level.

During the course of Oxera's assessment, the Commission's final proposals were published.[3] Most of the proposed rules are more or less in line with the original proposals and therefore have not affected Oxera's analysis.[4] The Commission's own impact assessment has been incorporated, as far as possible, into Oxera's analysis.

1.2 Approach and sources of information

This report applies a conceptual framework in which the Commission's proposals are assessed from a policy perspective. It systematically examines the potential justifications for regulation by assessing the extent to which concerns about the impact of HFT that have been raised in the literature could mean that there are specific market failures that need to be addressed. Furthermore, it assesses the impact of the proposed rules on the market and end-investors, and, importantly, the mechanisms through which the impact would arise.

In order to inform its understanding of HFT and the potential impact of the proposed rules, Oxera conducted interviews with market participants, academics who have specific expertise in the area of trading and the functioning of capital markets, and other stakeholders; and reviewed the literature on HFT.

The assessment presented in this report should be seen as a high-level impact assessment.[5] The scope of the research was restricted to the information sources listed below—any empirical analysis was beyond the scope of the report and would require more time and resources. It is not the purpose of this report to provide answers to all questions in relation to HFT. Many questions are likely to require detailed empirical analysis.

In the debate on HFT, there is some confusion and a diversity of views. To some extent, this may be due to people defining HFT in different ways (or not defining it) and not being clear about the particular concerns or market failures in relation to HFT. However, the nature of HFT also seems to be subject to considerable change, which may affect its impact, and the need for and impact of regulation. There also seems to be a difference between the nature and extent of HFT in the USA compared with Europe. For example, the ratio of messages to executed orders (in relation to HFT) seems to be higher in the USA than in Europe. HFT may have been developed more in the USA as a result of lower trading and post-trading transaction fees, and possibly as a result of the way in which different trading venues are linked together.[6]

The following information sources were used.

- *Academic literature*—there is a growing body of academic literature on the impact of HFT. The purpose of this economic impact assessment is to put the findings of academic research

[3] European Commission (2011), 'Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments repealing Directive 2004/39/EC of the European Parliament and the Council', European Commission, October 20th.

[4] As explained in section 4, the final proposal for a market-making requirement seems to go beyond the original proposal.

[5] Similar to high-level cost–benefit analyses undertaken by the Financial Services Authority.

[6] For example, Regulation National Market System (RegNMS) in the USA and the best-execution requirements on brokers in Europe. See, for example, Gomber, P., Arndt, B., Lutat, M. and Uhle, T. (2011), 'High-Frequency Paper', commissioned by Deutsche Börse.

into a policy-making context rather than summarising the literature. Useful reviews of the academic literature can be found in Gomber et al. (2011), Penalva (2010) and Biais and Woolley (2011).[7] A number of books have been written on automated trading and HFT in particular.[8]

- *Interviews with stakeholders*—Oxera conducted interviews with a large number of market participants (for example, fund management, hedge, brokerage firms and market-making firms, HFT traders), infrastructure providers, regulators, academics and other stakeholders. Oxera also participated in various seminars and workshops that were taking place. The interviews were used to obtain a better understanding of the rationale behind the trading strategies pursued by high-frequency traders, and the potential impact of some of the proposed rules.

- *Studies and consultations by regulatory authorities*—in addition to the rules proposed by the MiFID review, there are a number of other studies and consultations published by authorities (eg, a consultation paper by ESMA)[9], reports by national financial regulatory authorities in Europe (eg, a paper by the Netherlands Authority for Financial Markets)[10] and outside Europe (eg, studies by the SEC and CFTC in the USA),[11] and a consultation document by IOSCO.[12]

- *Responses to consultations and other papers by market participants*—a selection of responses to the MiFID review, the SEC/CFTC study and the IOSCO consultation were reviewed.[13] Some market participants have published their own position paper or presentation on the topic of HFT (eg, Optiver, Nanex and Tradeworx).[14]

This report does not provide a quantification of the impact of the proposed rules. Such quantification would require an empirical analysis, which would have necessitated more time and resources than were available.

During the course of the research, it became clear that a quantification is not necessarily required for a high-level impact assessment of the appropriateness of the rules. Applying economic logic allows for an assessment of how the proposed rules could affect the market. The fact that no quantification is provided (of the positive or negative impact) does not mean that the impact is likely to be small. On the contrary, any small changes in the regulation of capital markets can have significant consequences for the efficient functioning of these markets

[7] Gomber et al. (2011), op. cit. Penalva, J. (2011), 'High Frequency Trading: An Overview', July, available at: http://www.josepenalva.com. Biais, B. and Woolley, P. (2011), 'High Frequency Trading', Working paper IDEI, March.

[8] See, for example. Narang, R. (2009), *Inside the Black Box: The Simple Truth about Quantitative Trading*, Wiley Finance, September. Aldridge, I. (2010), *High Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*, Wiley Trading. Perez, E. (2011), *The Speed Traders: An Insider's Look at the New High-Frequency Trading Phenomenon that is Transforming the Investment World,* McGraw-Hill, April.

[9] ESMA (2011), 'Consultation Paper: Guidelines on Systems and Controls in a Highly Automated Trading Environment for Trading Platforms, Investment Firms and Competent Authorities', Consultation Paper, European Securities and Markets Authority, July 20th.

[10] AFM (2010), 'High Frequency Trading: the application of advanced trading technology in the European market place', report, the Netherlands Authority for Financial Markets.

[11] SEC/CFTC (2010), 'Findings regarding the market events of May 6, 2010', report, US Securities & Exchange Committee and US Commodity Futures Trading Commission, September 30th.

[12] IOSCO (2011), 'Regulatory Issues Raised by the Impact of Technological Changes on Market Integrity and Efficiency', Consultation report, International Organization of Securities Commissions, July.

[13] At the time of producing the report, responses to the ESMA consultation were not yet available.

[14] Optiver (2010), 'High Frequency Trading', Position Paper, December, www.optiver.com/corporate/hft.pdf. Tradeworx (2010), 'Public Commentary of SEC Market Structure Concept Release', April, available at: www.tradeworx.com/TWX-SEC-2010.pdf. Nanex (2010) 'Nanex Flash Crash Summary Report', September 27th, available at: www.nanex.net/FlashCrashFinal/FlashCrashSummary.html.

and may affect end-users such as pension fund holders and firms that raise capital, and, ultimately, may affect the real economy.

1.3 Main conclusions

Market-making requirement

The MiFID consultation paper proposed that:

> Market operators would be required to ensure that if a high frequency trader executes significant numbers of trades in financial instruments on the market then it would continue providing liquidity in that financial instrument on an ongoing basis subject to similar conditions that apply to market makers;[15]

In the Commission's final version of its proposals, this has been reworded as follows:

> An algorithmic trading strategy shall be in continuous operation during the trading hours of the trading venue to which it sends orders or through the systems of which it executes transactions. The trading parameters or limits of an algorithmic trading strategy shall ensure that the strategy posts firm quotes at competitive prices with the result of providing liquidity on a regular and ongoing basis to these trading venues at all times, regardless of prevailing market conditions.[16]

One of the differences between the original proposal and the final version is that there is no longer a reference to the fact that these high-frequency traders would be subject to 'similar conditions that apply to market makers'.

According to the Commission, the rationale behind the proposed rule is to ensure that high-frequency traders provide meaningful liquidity at all times and would contribute to more orderly and liquid markets and mitigate episodes of high uncertainty and volatility. In other words, the ultimate aim is to prevent market panics and crashes from happening.

It is useful to bear in mind that market-making requirements were never introduced with the intention to prevent panics or market crashes, nor to maintain artificially the old price of a security once additional information on that security has become available and it now has a new equilibrium fair price. They were introduced with the aim of providing immediacy—ie, ensuring that a buyer in the morning does not need to wait until the afternoon to find a seller.

The way the proposal is now worded, however, means that all AT will be required to use an algorithm that 'posts firm quotes at competitive prices ... at all times'. Although all HFT can be seen as being algorithmic, not all AT is high-frequency. Discussion with market participants suggests that there is very little trading that is purely manual. Agency orders that result in long-term changes in net positions are often executed using a (non-high-frequency) algorithm. Non-high-frequency market-making may also use algorithms to maintain the market-makers' market position, rather than relying on direct human intervention to update those positions as their resting orders execute and additional information is incorporated into the security's price. As a result, the majority of market participants would appear to be caught by this requirement to post firm quotes at competitive prices at all times. The impact of such a requirement would,

[15] European Commission (2010), 'Review of MiFID', December 8th, p. 16.

[16] European Commission (2011), 'Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p. 70.

therefore, extend significantly beyond high-frequency traders. This does not seem to have been taken into account in the Commission's impact assessment.

The impact of this market-making requirement will depend critically on how the requirement to 'continually post firm quotes at competitive prices' is interpreted. At one extreme, there will be significant additional risk applied to AT if the requirement is interpreted to mean that at all times the market is open, and that every algorithmic trader has to offer to buy and sell a security across a spread that reflects the usual spread for that security. At times of increased uncertainty those required to maintain bid and offers across the usual spread are likely to find that they enter into a significant number of trades where the subsequent track of prices will have moved against them and which they will not be able to unwind without loss.

This could have the effect of making AT (including buy-side trading using computer-controlled execution) non-viable because of the requirement to provide firm bid and offers at competitive prices being too risky. All trading could, therefore, return to manual trading. If, as seems likely, the return to manual trading increases the total costs of intermediaries (because they have to substitute people for computers) this increase in intermediaries' costs will be paid for by investors (which will see a lower net return) or companies (which will see an increase in the cost of capital which will then be passed on in the form of higher prices in the end product market).

If, on the other hand, the interpretation of the requirement is designed to mimic the requirements applied currently to official market-makers, so that there is still the possibility of market-making using computer-driven trading sequences being profitable, then, the impact on HFT sequences is likely to be minimal.

In general, this arises because any HFT sequence algorithm is likely to be able to have a market-making module bolted on to it that will *display* firm bid and offers at competitive prices under normal market conditions, and to widen the spread offered or withdraw from the market under those market conditions where bid and offers across a narrow spread become highly risky (as traditional market-makers are generally currently allowed to do). Such a module could be designed to actually execute very rarely, if at all, by the repositioning of these (resting) orders before they reach the head of the queue. In addition, if the definition of 'competitive' price includes prices one or more ticks away from the touch, then ensuring that actual execution is limited, will be easier. The re-positioning of orders (before they search the head of the queue) could result in an increase in the number cancellations relative to executions.

Unlike those intermediaries which wish to make money out of market-making, those which simply wish to meet this requirement in order to pursue other strategies will generally only wish to avoid losing money, and therefore will not be concerned with executing trades, at some point, for the purpose of market-making. The specific implications for the different trading sequences are analysed in section 4.5.

In sum, the impact of the proposed market-making rule will depend on the precise interpretation. In one scenario, the proposed rule may have a limited impact but at the same time is unlikely to achieve the Commission's objective of improving liquidity and reducing volatility. In another scenario, the proposed rule would have the potentially severely negative unintended consequence of all trading returning to manual trading resulting in an increase in costs of intermediaries.

Minimum tick size

The MiFID review proposes that implementing measures could further specify minimum tick sizes—it does not prescribe specific minimum tick sizes at this stage. Needless to say, the minimum tick sizes would be applied to all trading and not simply automated trading or HFT.

The MiFID review consultation document does not explain the objective of this proposed rule, so it is not possible to analyse directly what market failure this proposed regulation is aimed at, nor is it therefore possible to directly evaluate if an alternative intervention would be available. However, given the form of the rule (minimum tick size) the general objective would seem to be that in the absence of the rule tick sizes would become (or currently are) 'too small'. If the rule is to have an impact it would seem to need to either:

- increase the tick size that would otherwise occur; or (as a side effect)
- coordinate and standardise tick sizes for the same security trading in different venues.

It is possible that the Commission considers that there is a need to impose a minimum tick size in Europe if it considers that tick sizes have become (or could become) too small, and that this small size itself causes direct harm to the capital market. However, given the context of this proposal, it is also possible that the proposed rule is aimed more generally at curtailing HFT, or at least certain trading strategies that are implemented using HFT.

Oxera's analysis suggests that limited increases in the minimum tick size would be unlikely to have a significant impact on either the volume or location of (high-frequency) trading. At the margin, some HFT sequences will become slightly more profitable, but may occur slightly less frequently. For example, in the case of arbitrage strategies, an increase in tick size will tend to increase the value of fleeting mispricing as the price in each venue momentarily moves apart. Where the mispricing is one tick (for example, when the price in the leader venue has moved, but the price in the follower venue has not) a bigger tick will increase the value of that mispricing (for any given volume of security). However, the slightly larger tick may also slightly reduce the frequency with which these pricing anomalies occur.

There is some empirical evidence (referred to in section 5) that very large increases in tick size could move trading away from lit venues.

Minimum resting times

The MiFID consultation paper proposed that:

> Market operators would be required to ensure that orders would rest on an order book for a minimum period before being cancelled.[17]

In the Commission's near final version of its proposals, this had been reworded as follows:

> Impose minimum latency period of orders in the order book—Under this option an obligation would be implemented according to which orders on electronic platforms would need to rest on an order book for a minimum period of time before they can be withdrawn.[18]

[17] European Commission (2010), 'Review of MiFID', December 8th, p. 16.

In the final set of proposals published by the Commission this requirement had been dropped. However, since there is still some degree of fluidity in the final set of interventions, the analysis that follows below may still be relevant.

According to the Commission, the proposed rule would stop high-frequency traders and algorithmic traders from testing the depth of order books by submitting and cancelling orders in very quick succession. This would put less stress on the IT systems of market operators, reducing the risk of system failures.

The Commission mentions that the proposed rule has certain disadvantages and concludes that the costs are likely to outweigh its benefits. This rule is therefore not one of the Commission's preferred options. It prefers to impose an order to executed transactions ratio.

A minimum resting period will bite only to the extent that the interval between an order arriving at the back of the queue and it being available for execution at the front of the queue is less than the resting period.[19] Much more detailed empirical analysis of the distribution of this time interval would be necessary to gain an understanding of the relationship between the value chosen as the minimum resting period and the impact on trading sequences, and hence the impact on HFT. In addition, an analysis of the predictability of this time interval in relation to the placing of an individual resting order would also be necessary to capture the likely response of HF traders.

If the individual time intervals (ie the time taken between arrival at the back of the queue at the touch and reaching the head of the queue) are highly predictable HF traders can be expected to continue to place orders at the back of the queue where the prediction indicates a high probability of the time interval being larger than the minimum resting period, and not to place orders at the back of the queue at the touch when the prediction is that the time interval will be less than the minimum resting time. In addition, where the time interval is not very predictable orders again will tend not be placed at the back of the queue at the touch.

Placing orders at the back of the queue in ticks away from the touch will, by definition, place more orders ahead of the relevant order and, therefore, increase the probability of the time to possible execution interval being longer than the minimum resting period.

As a result, HF traders can be expected to reduce the placing of orders at the back of queue in the tick at the touch, and to increase the placing of orders in ticks away from the touch (or what would have been the touch in the absence of the minimum resting period rule). In addition, in those periods of the trading day where there is a relatively rapid execution of orders this behaviour response by HF traders can be expected to be more pronounced.

Generally, therefore, as the minimum resting period rule bites (ie as it is made longer) HF traders would tend to place fewer (if any) orders at the touch (or what would have been the touch) and to potentially place more pre-positioning orders away from the touch. Observed spreads would, therefore, tend to widen.

Although the relationship between widened spreads and the price of immediacy is not straightforward (especially for large investor orders), they will tend to translate into a higher

[18] European Commission (2011), 'Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p. 70.

[19]

price for immediacy, particularly for smaller investor orders. Because of the volume of trading, a small increase in the price of immediacy can have a large total impact. In order of magnitude terms, if the (approximate) market capitalisation of the FTSE 100 (£1.5tr) is taken as the proxy for the market capitalisation of all European equity securities being currently traded in a significant way by HF traders, and the investor turnover of those securities purchased or sold with immediacy is taken as 0.5, a one tick increase (~5bp) in the effective spread would result in an increase in the cost of immediacy of around £375m. **It is difficult to exactly predict the impact of an increase in resting times on the bid-ask spread but this analysis suggests that if the increase in resting times were large, the impact in terms of the increase in the costs of immediacy could be billions rather than millions.**

Increasing the minimum resting time is also likely to have a significant impact on the use by (high-frequency) traders of orders designed to discover (additional) information with respect to current market (order book)—ie, pinging orders. Pinging refers to the practice of sending an order to a trading venue to try to discover if there is demand for that security (either buy or sell) that is not otherwise visible. Ping orders being used to provide information for use in market-making, directional or arbitrage trading sequences will tend, therefore, to be such that if they do not execute on arrival (ie, identify the hidden liquidity) the high-frequency trader will not want them to execute. However, even very short minimum resting periods will be likely to increase the probability of this happening, and since these executions will generally be at the wrong time, it would be expected that the use of pinging orders would decrease markedly.

The reduction in the information available to HFT will in turn have some impact on the ability to make predictions on the future direction of prices, so that some reduction in trading sequence activity could be expected. However, the extent to which the information generated by pinging creates *relative* advantages between high-frequency traders, rather than an absolute ability to make short-term price predictions, the reduction in non-pinging trading sequences will be muted.

Pinging is also used by non-HF traders (eg long only investors) wishing to minimise the market impact of a large change in a net position. By identifying where there is, or is not, hidden liquidity, the trader can attempt to optimise the order placements at different venues so as to minimise market impact. However, in these cases, it is likely that the execution of a ping order as a result of a minimum resting requirement would represent a transaction that is actually wanted and would be executed at a better price than the visible orders at that time. A minimum resting period would be likely to have much less impact on the use of pinging orders for this type of activity.

In sum, imposing minimum resting times is likely to have a significant impact on pinging. To assess the ultimate impact on the market, three questions need to be answered.

- First, what proportion of trading messages (including cancellations) is driven by pinging? Although in principle, it would be possible to analyse this, currently there is no readily available data analysis that has broken down the messaging patterns to the type of HFT sequence.

- Second, to what extent do other HFT strategies (such as directional strategies) rely on the absolute amount of information obtained by pinging and could these strategies still be implemented based on information from other sources?

- Third, more generally, what is the value of pinging to the market and could there be any negative consequences arising as a result of a reduction in the amount of pinging?

Minimum execution ratios

The MiFID consultation paper proposed an alternative to the minimum resting time rule:

> .... they would be required to ensure that the ratio of orders to transactions executed by any given participant would not exceed a specified level. In either case, further specification would be needed on the specific period or level.[20]

In the Commission's final version of its proposals, this has been worded as follows:

> Member States shall require a regulated market to have in place effective systems, procedures and arrangements to ensure that algorithmic trading systems cannot create or contribute to disorderly trading conditions on the market including systems to limit the ratio of unexecuted orders to transactions that may be entered into the system by a member or participant, to be able to slow down the flow of orders if there is a risk of its system capacity being reached and to limit the minimum tick size that may be executed on the market.[21]

The Commission concludes that the order to execution ratio rule can achieve the same objectives as imposing a minimum resting time (ie, it would stop high-frequency traders and algorithmic traders from testing the depth of order books by submitting and cancelling orders in very quick succession, thereby putting less stress on the IT systems of market operators and reducing the risk of systemic failures), but that the disadvantages would be less severe.[22]

Oxera's analysis suggests that it should be possible for high-frequency traders to modify their order flow to venues in ways that would meet the required ratio, but without having a really significant impact on the orders that do, actually, execute. Since the trading sequences are already calculating probabilities of execution, the most likely outcome is that marginal order sequences (ie, those with predicted very low probabilities of resulting in an execution) would be curtailed, but those with a reasonably high probability would continue (unless the ratios are set very tightly). It may also be possible for high-frequency traders to 'create' executing trades if required. Although such an approach would incur the trading costs of using the relevant venue, it could be carried out without any market impact or loss across the spread because the trader would be on both sides of the trade.

It should be borne in mind that the ratio of unexecuted orders to transactions will vary considerable according to the asset being traded, with typically much higher ratios in those asset classes with strong pricing relationships with other assets, especially where the asset trades infrequently. This is a function of how the assets are priced and how risk is distributed. Accordingly it would be inefficient and possibly dangerous to try and impose a single ratio on all trading, but rather any restrictions should be set with reference to the specific pricing relationship and process. For example, Apple as a price leader would have a much lower unexecuted order to execution ratio compared to an illiquid stock that is priced with reference to the Apple price, but only trades once a day. Similarly, a corporate bond that maybe trades once a quarter would have a much higher ratio that the government bonds it is priced with

[20] European Commission (2010), 'Review of MiFID', December 8th, p. 16.

[21] European Commission (2011), 'Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p. 116.

[22] European Commission (2011), 'Commission staff working paper, Impact Assessement accompanying the document: Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p.129

reference to. These differences will also apply at the level of the asset class so equities as a class will have a very different ratio to bonds etc.

As a result of these considerations, and because high-frequency traders are likely to use probabilistic approaches to meet a requirement of this sort, any impact on HFT would be at the margins (unless the ratio is set very tightly). In addition, a rule of this sort could have the unintended consequence of artificially favouring larger and multi-client type market participants over smaller and specialist operations that may find it more difficult to meet the order-execution.

Circuit breakers

As circuit breakers (or equivalent) rules are already largely in place, the impact of the proposed rules here is limited. This is analysed in more detail in section 8.

1.4 Structure of the report

This report is structured as follows.

- Section 2 sets out Oxera's approach to assessing the impact of the proposed rules.
- Section 3 outlines the primary generic trading sequences used by high-frequency traders.
- Sections 4 to 8 present the impact of the proposed rules on the sequences described in section 3, along with the more general impact of the proposed rules.
- Section 9 provides a high-level impact assessment of further increases in the speed of the operation of the proposed rules.

2 Approach to assessing the impact of the proposed rules

2.1 Methodology

An impact assessment normally consists of two steps: first, identifying and assessing the market failure(s) that the proposed regulation intends to address; second, assessing the impact of the proposed rules in terms of the costs and benefits. Existing methodologies typically refer to the following types of costs and benefits.[23]

- *Direct costs*—designing, monitoring and enforcing regulations requires resources. The value of the extra resources that would be absorbed by the regulatory regime in respect of a proposal is the direct cost of that proposal.

[23] Existing methodologies and frameworks for impact assessment include: European Commission (2009), 'Impact Assessment Guidelines, European Commission, January 15th; Oxera (2006), 'A framework for assessing the benefits of financial regulation', report prepared for the Financial Services Authority, September; FSA (2006), 'A Guide to Market Failure Analysis and High Level Cost Benefit Analysis'; HM Treasury (2003), 'The Green Book, Appraisal and Evaluation in Central Government', (and subsequent amendments).

- *Compliance costs*—the value of extra resources (including time) that would be used by firms and/or individuals to comply with a regulatory proposal is known as the compliance cost of that proposal. This refers to administrative costs, the costs of putting in place new systems and procedures to implement the regulation, and monitoring on-going compliance.

- *Indirect costs*—the indirect costs refer to negative market impacts, for example, reduced competition or reduction in the quantity and quality of the goods produced or offered. In other words, the regulation may have negative effects on market outcomes which, in the case of regulation of trading, can refer to price formation and market efficiency.

  Generally speaking, indirect costs are harder to measure than direct and compliance costs but can also be assessed (to a large extent) qualitatively. In practice, since indirect costs refer to how the market operates they can potentially be very significant and are more likely to drive the outcome of an impact assessment than the direct and compliance costs. They are referred to as indirect rather than direct costs since they arise indirectly—for example, through the way in which firms (those that are subject and/or potentially those that are not subject to the regulation) and consumers respond to the regulation, or through the way in which firms interact following the implementation of the regulation.

  Indirect costs may arise in various ways. For example, the regulation may partly achieve its objective (by addressing the market failure), but because of the way the rule has been specified, it may affect certain firms more than others and potentially distort competition. It may also be the case that the regulation is too blunt in that it bans or curtails a group of activities, some of which may be harmful, whereas others may be beneficial. Regulation that is more targeted is therefore less likely to result in indirect costs.

  Indirect costs can be assessed by identifying the mechanisms through which market outcomes can be negatively affected. Firms and consumers may respond in numerous ways, and their activities and services may be affected in numerous ways.[24]

- *Benefits*—the assessment of benefits can, to some extent, be considered the ‘flip side’ of the market failure analysis. If there is a market failure and the regulation is effective, then there are likely to be benefits, equivalent to an improvement in market outcomes, as a result of addressing the market failure. In other words, benefits are probably only going to be significant if the market failure was indeed a significant concern. If there is no clear justification for imposing regulation, then it may also be more difficult to identify benefits. Benefits often arise indirectly rather than directly (except, for example, if a certain product or practice is banned that is harmful in itself). This means that, similar to assessing indirect costs, the assessment of benefits requires an identification of the mechanisms through which they improve market outcomes.

2.1.1 Practical implications for Oxera’s assessment

This impact assessment framework has been applied to the Commission’s proposed rules in relation to HFT in the following way.

[24] For a more detailed explanation, see Oxera (2006), ‘A framework for assessing the benefits of financial regulation’, report prepared for the Financial Services Authority.

- *Market failure analysis*—Foresight has not asked Oxera to undertake a market failure analysis. The impact of HFT in terms of positive and/or negative effects on the market has been assessed in other papers commissioned by Foresight, other academic articles in the literature and, to some extent, in policy papers such as the consultation paper published by IOSCO.

- *Focus on indirect costs and benefits*—since this is a high-level impact assessment, an assessment of the direct and compliance costs is beyond the scope of this study. Although not necessarily insignificant, the direct and compliance costs of the proposed rules are unlikely to drive the outcome of an impact assessment. If there are any significant costs, these would be more likely the result of a negative impact on the market. Hence the analysis in this report focuses on assessing the indirect costs and the benefits.

As explained in more detail below, HFT is used to implement a number of different trading strategies. These trading strategies are analysed and described in terms of trading sequences in section 3. Sections 4 to 8 then assess the impact of the proposed rules by looking at how their application would affect the trading sequences outlined in section 3. If the rule has an impact, it is also possible to evaluate if there are any (relatively simple) changes to the sequences that would tend to change (and potentially eliminate) the impact of the rule. In this way it is possible to build up an evaluation of the likely medium-term impact of the rule on market dynamics and to evaluate the further impact of increasing the speed at which high-frequency traders can access information from, and send information to, trading venues.

2.2 What is high-frequency trading?

HFT is a subset of AT, which uses computers to handle order flows to trading venues, rather than handling those flows manually.[25] It is not a trading strategy in itself, but a tool or technique to implement certain trading strategies. There is no single agreed definition of HFT—various studies have attempted to provide one. Examples include:

> [HFT is trading that] uses sophisticated technology to try to interpret signals from the market and, in response, executes high volume, automated trading strategies, usually either quasi market making or arbitraging, within very short time horizons ... [and] involves execution of trades as principal (rather than for a client) and involves positions being closed out at the end of the day.26
>
> HFT entail strategies trading with investment horizons of less than one day, and seeking to unwind all positions before the end of each trading day27
>
> High frequency trading (HFT) is a method of implementing certain short-term trading strategies using advanced technology, but is not in itself a separate trading strategy.28
>
> HFTs are proprietary trading firms that use high speed systems to monitor market data and submit large numbers of orders to the markets. HFTs utilize quantitative and algorithmic methodologies to maximize the speed of their market access and trading strategies. Some HFTs are hybrids, acting as both proprietary traders and as market-makers. In addition, some HFT strategies may take 'delta-

[25] See Cartea, Á. and Penalva, J. (2011), 'Where is the Value in High Frequency Trading?'. The authors define algorithmic trading (AT) as a generic term that refers to strategies that use computers to automate trading decisions, and restrict the term 'high frequency (HF) trading' to a subset of AT trading strategies that are characterised by their reliance on speed differences relative to other traders to make profits based on short-term predictions, and also by the objective to hold essentially no asset inventories for more than a very short period of time.

[26] European Commission (2010), 'Review of MiFID', December 8th, p. 14.

[27] Tradeworx (2010), 'Public Commentary on SEC Market Structure Concept Release', April 21st, p. 4.

[28] AFM (2010), 'High frequency trading: The application of advanced trading technology in the European marketplace', November, p. 5.

> neutral' approaches to the market (ending each trading day in a flat position), while others are not delta-neutral and sometimes acquire net long and net short positions.29

> [t]rading activities that employ sophisticated, algorithmic technologies to interpret signals from the market and, in response, implement trading strategies that generally involve the high frequency generation of orders and a low latency transmission of these orders to the market. Related trading strategies mostly consist of either quasi market making or arbitraging within very short time horizons. They usually involve the execution of trades on own account (rather than for a clients) and positions usually being closed out at the end of the day.[30]

IOSCO identifies the following common features and trading characteristics related to HFT:[31]

> it involves the use of sophisticated technological tools for pursuing a number of different strategies, ranging from market making to arbitrage;
>
> it is a highly quantitative tool that employs algorithms along the whole investment chain: analysis of market data, deployment of appropriate trading strategies, minimisation of trading costs and execution of trades;
>
> it is characterized by a high daily portfolio turnover and order to trade ratio (ie, a large number of orders are cancelled in comparison to trades executed);
>
> it usually involves flat or near flat positions at the end of the trading day, meaning that little or no risk is carried overnight, with obvious savings on the cost of capital associated with margined positions. Positions are often held for as little as seconds or even fractions of a second;
>
> it is mostly employed by proprietary trading firms or desks; and
>
> it is latency sensitive. The implementation and execution of successful HFT strategies depend crucially on the ability to be faster than competitors and to take advantage of services such as direct electronic access and co-location.

An additional characteristic is that the trading strategies are typically informed by an analysis of order book data rather than information that would directly reveal something about the fundamentals of the (underlying) security.

Given the variability of the existing definitions, it may be more pragmatic to define HFT with respect to strategies for profitable trading that can now be exploited as a result of changes in the microstructure of markets. Such an approach is set out below.

2.2.1 What drives high-frequency trading?

HFT makes it possible to implement a specific set of trading strategies, which can potentially be defined as the generation of profit from buying and selling the same security(s) over very short time periods by predicting (on a probabilistic basis) movements in prices (including fleeting pricing anomalies between trading venues or between linked securities) or by executing a transaction across a spread, or both. The following conditions are essential to delivering successful implementation of such trading strategies.

[29] US Securities & Exchange Committee and US Commodity Futures Trading Commission (2010), 'Findings regarding the market events of May 6, 2010', report, September 30th, p. 45.

[30] ESMA (2011), 'Guidelines on systems and controls in a highly automated trading environment for trading platforms, investment firms and competent authorities', July 20th, p. 10.

[31] IOSCO (2011), op. cit.

- The ability to complete quickly the transaction round trip (ie, the buy and the sell sequence) across the spread must be predictable—at least on a probabilistic basis.

- Price changes must be predictable—at least on a probabilistic basis.

- Absolute speed—for the price changes to be predictable, the description of the market conditions and the time at which an order can be executed based on an understanding of those market conditions must be close in terms of minimising the probability of other events occurring (including trading by other market participants) that may alter those market conditions.

- Relative speed—given the limited supply of securities at the right price, and the fact that the execution of orders alters the market conditions, traders have to be first (ie, faster) than other traders, and transaction costs must be lower than the price differential that is exploited.

- The total cost of the capital employed in the transaction must be low (ie, the same capital must be used very frequently).

- The prices for the specific transaction services need to be low (ie, trading and post-trading fees).

- The cost of capital and labour per unit of transaction must also be low. Put the other way round, the capital equipment and labour costs need to be spread over a large number of transactions occurring in a short space of time.

This framework can help explain the emergence of high-frequency traders, as follows:

- market fragmentation increases the predictability or probability of certain price changes occurring—eg, instantaneous price differentials in different trading venues;

- the development of derivatives and indexes markets creates the conditions for more predictable sorts of price change—in particular, the price changes that bring relative prices back to their ‘normal’ relationships;

- transaction costs (trading and post-trading fees, etc) have been falling. In addition, some trading venues have created pricing structures that allow certain transaction sequences to be charged very low transaction fees, or, in some cases, for the price to the trader to be negative (ie, the exchange will pay or rebate traders executing through particular types of order);

- speed of data analysis and the ability to handle/analyse very large amounts of data have improved and become cheaper, thus making the ability to predict either round-trip times across the spread, or short-term price movement, both economic and within the relevant time horizon;

- because a trader’s speed relative to others wishing to exploit the same profitable trading opportunities is important, there is competitive pressure to continually increase the overall speed of these trading strategies and sequences.

Section 3 describes, in generic terms, the trading sequences that lie behind these trading strategies.

2.2.2 High-frequency trading and intermediation, or how high-frequency trading can be profitable

Trading (as opposed to investment) constitutes transacting with counterparties in the absence of any fundamental long- or medium-term investment opinion (and as a result in absence of any opinion on the long-term direction of the price) in respect to the underlying asset being traded. It essentially relies on an assessment of the short-term supply and demand conditions of the asset being traded. In isolation, trading can earn a return (ie, can be profitable) by supplying services to investors who, if trading is to be profitable as an activity, provide the other side (ie, the counterparty) to the traders' transactions.[32] Traders as a group can make money from investors by providing liquidity to a marketplace which facilitates transactions and providing a service of immediacy and intermediation.

HFT is nothing more than the use of technology to perform this function very quickly and, from the high-frequency traders' perspective, frequently, and to provide very quick interlinkages between markets in ways that reflect their structural or stochastic connections as a means of deriving value and hedging correlated risk. It should also be noted that, in general, traders can only remain in business by actually trading. Particularly in relation to HFT, where there are very high rates of order cancellation. The profitable activity is based on actually undertaking transactions, not on just placing and cancelling orders.

All trading, including HFT, will tend to be based on one or more of three fundamental types of intermediation: of time, of place and of form.

Intermediation of time—or the service of immediacy

This is the provision of the service of immediately providing the counterparty to an order placed by an investor, and is undertaken by offering an immediately executable price (to buy or to sell a specific quantity) in an asset in between the points in time where other parties who have a fundamental economic interest in buying or selling the asset express their offsetting appetite through their own placement of orders to buy or sell.

When this service is provided the trader takes on some form of market risk, potentially mitigated by forms of risk transfer or elimination derived from the other two forms of intermediation described below (place and form). However, since risk transfer or mitigation can never be achieved instantaneously, some level of market risk is always involved. The extent of that risk depends on the VaR[33] associated with the risks incurred throughout the process, which is itself derived from the properties of the risks (such as volatility, extent of mean reversion,[34] properties of the risk distribution, and adverse selection[35]) and the period of time for which those risks are held.

[32] If traders only trade with themselves the activity of trading is a zero-sum game, so overall the activity cannot be profitable.

[33] Value at risk: a technique used to estimate the probability of portfolio losses based on the statistical analysis of historical price trends and volatilities.

[34] A theory suggesting that prices and returns eventually move back towards the mean or average. This mean or average can be the historical average of the price or return, or another relevant average such as the growth in the economy or the average return of an industry.

[35] A trader providing the service of immediacy is at risk of trading with counterparties with more information (for example, that a firm is about to issue a profits warning) and, therefore, their trading partners' order flow is not random (in this case, there would be more sell orders).

The less liquid an asset, the less likelihood there is of investors' offsetting interests coinciding in time. The intermediating trader therefore provides more value and typically earns higher economic rent per transaction for less liquid assets.

For example, if there are a handful of investors who express a fundamental interest in buying or selling a stock no more than twice a year, the chance of them having offsetting interests at the same time is very close to zero. The intermediation service performed by the trader is therefore highly valuable, and the extent of the risk incurred in doing so high, since it is a function of the time the trader is exposed to the danger that the fundamental price of the security moves against the trader. At the opposite extreme, if an asset is traded thousands of times per second by multiple market participants,[36] the probability of them having offsetting interests at the same time is significantly higher, so the trader offering intermediation of time (immediacy) is therefore less valuable and the economic rent (reward) for providing the immediacy service is low. Similarly, the intermediation function only involves holding risk for a very short period of time, so the cost of providing the service is also low for this reason and can be priced accordingly (ie, low).

The price of the service of immediacy can be thought of as the spread between an offer to buy immediately (eg, at £10.00) and an offer to sell immediately (eg, at £10.10). The costs of the service are the cost of using capital to hold the security until both transactions complete, and the risks (or benefits) are that the price that can be secured for the second transaction moves against (in favour) of the trader. Reducing the holding time and reducing the holding risks (or increasing the probability of the price moving in the favourable direction) reduces the costs, and hence the spread (ie, price), at which the provision of the service is economic.

Many assets exhibit a high degree of structural dependency, or statistical correlation, to other assets, and the ability of the risk to be hedged with a high degree of precision by reference to these as sources of pricing information or risk mitigation/elimination is also fundamental to the costs (economic rent) of intermediation.

The extent of that structural dependency or correlation is therefore pertinent, and whether it represents structural dependency or just time-series correlation has profoundly significant implications for the type and degree of risk, and the consequent equilibrium level of costs (economic rent) associated with the provision of the service of immediacy.

Intermediation of time is therefore fundamentally dependent on, and connected to, the other two forms of intermediation described below.

Intermediation of place—or keeping prices of the same security the same when it is traded in more than one place

This is the provision of liquidity in an execution venue physically isolated from other execution venues where the same asset (or assets whose prices are defined with reference to the same sets of underlying economic valuation parameters and risks) is traded.

The trader provides the service of ensuring that the price for the same asset is the same in different venues, or, more specifically, is only different to the extent that there are real costs

[36] It should be noted that even very liquid securities rarely trade at this frequency (see section 6).

(eg, transport costs, insurance, risk pricing) involved in its movement from one venue to another.

Whereas this function is relatively obvious in the context of physical commodities, it is still pertinent in relation to dematerialised assets, such as financial instruments, to the extent that there are real costs associated with their ‘logical’ transit between execution venues. Even in cash-settled markets those costs are not zero due to various operational criteria, for example, technology costs and the price of the risk consumed during the transit process. The economic rent earned by the trader providing intermediation of place is therefore dependent on the optimisation of those transit costs.

Intermediation of form—or keeping the price of similar or linked securities similar

The fundamental value of different assets is linked through the workings of the economy and, as a result, assets are generally priced with reference to their expected relationship (correlation) with other assets. These price relationships may be:

- *definitive*—where the instruments are structurally identical, ie, correlation is 100%;
- *highly structurally dependent*—where the instruments are fundamentally related in terms of their pricing, but different to some extent;
- *purely stochastic*—where there is no fundamental structural relationship between the assets, but their prices are observed to behave in a correlated way.

A trading strategy based on a *definitive* relationship rarely works (at least on its own) in competitive markets because it is so generic, and can be undertaken with such high certainty, that there is little cost and therefore no price for the service—the economic rent is arbitraged away. Therefore the opportunities for traders to earn economic rent from this tend to be extremely limited.

Trading strategies based on s*tructurally dependent* and *stochastic* relationships are much more common because new ways of capturing relationships are constantly being discovered (through historical data analysis) and exploited. In reality, most assets that are not definitively related to each other have a (price correlation) relationship that exists somewhere along a continuum of correlation between structural and stochastic, so these relationships are each very much a question of degree.

Additionally, arbitrage-free asset price relationships are often derived from a composite of pricing parameters, some of which may be structural and some of which may be stochastic. The relative weight of each of these pricing parameters within the overall pricing ‘recipe’ will determine both the extent to which the prices are correlated and the extent to which end of the continuum (structurally to stochastically) they are correlated.

In effect, with structurally dependent and stochastic relationships, there is an almost infinite web of variables in the potential pricing recipes used by different market protagonists, far too many to say that the price of A depends purely on the price of B. Because people do things for different reasons, the set of interests meeting concurrently in the marketplace is diverse, leading to varying valuations and prices for the same instruments. This represents the essence of why transaction volumes tend to be high relative to end-user investment transactions: changes in the multitude of input variables being used as pricing inputs and risk-management methods by different market participants cause their views on what is the right price to change. This results in transactions between them (ie, between traders) as they seek to ensure that their own evaluation of their risk remains within parameters consistent with their own underlying assumptions and to reduce their risks as a result of their own current market position at any one time.

These factors can lead to traders trading with each other, which, although a zero-sum game in terms of profits, may not be a zero-sum game in terms of costs. For example, and in its simplest form, two high-frequency traders may find themselves holding opposite sides of the same transaction (eg, one has sold a security and therefore needs to buy it to return to a flat position, while the other has bought the same security and needs to sell it to return to a flat position). Although between them they cannot make money from trading between themselves, by undertaking the transaction both of them can eliminate their market risk position, and so mutually reduce the *costs* of their trading. Where dependent or stochastic relationships are involved, the mutually beneficial trading between high-frequency traders is more complex, but the underlying rationale and impact (ie, on the cost side, not the income side) is the same.

Such processes and interactions between high-frequency traders have an operational dimension in that a fast and predictable outcome will minimise the extent of risk carried during the ‘manufacturing process’, and the potential loss associated with ‘fat tails’.[37]

In practice, most mature HFT strategies employ all of these relationships between securities to some extent. Although there does not appear to be any explicit analysis of the degree to which these different forms of relationships are used, industry discussions suggest that the relationships that underpin the intermediation of form are usually a fundamental part of HFT strategies, but that these are usually deployed in combination with intermediation of place as an optimisation technique. Traders’ overall objective function is to contribute passive or aggressive liquidity into intermediation of time (ie, the provision of the service of immediacy) as their output, in order to earn the associated value of this service as supplied to investors (ie, its economic rent).

Strategies that are based on only one of these types of relationship tend to become obsolete quite quickly, since obvious arbitrage returns are rapidly undermined by competitive forces in a highly dynamic marketplace with relatively low barriers to entry (at least in terms of the tangible assets required).[38] Generally, there is a profusion of new strategies being developed and tested by high-frequency traders, only a small proportion of which ultimately prove profitable, and whose duration of profitability may be limited. Durable strategies tend to require competitive performance close to the efficient frontier, in terms of production costs, and require multiple concurrent approaches in order to be sustainable, and to address some fundamentally useful role in creating liquidity for investors in order to earn an economic return.

2.2.3 The role of hedging and risk transfer within (high-frequency) intermediation

When the high-frequency trader first provides an immediacy (or other) service for an investor, the initial transaction leaves the high-frequency trader with an exposed (risk) market position. If

[37] A fat-tailed distribution is a probability distribution where extreme events, such as a large move up and/or down in prices, are more likely than given by a normal distribution.

[38] As there has been only a limited period of HFT activity, it is unclear if there are other significant barriers to entry. If the experience of HFT using very complex algorithms and analytics turns out to be of critical importance to success, there may be advantages to having entered this activity early which could create barriers to subsequent entry.

the trader can immediately (or very quickly) enter into a transaction that reverses the original transaction, the risk is completely eliminated. In many cases the short holding periods that characterise HFT activity can be seen as completing this round trip very quickly. However, if the transaction that reverses the original transaction is not immediately available, the risk exposure of the first transaction can be reduced by other transactions that may be available and that hedge, or in some other way reduce, the risk faced by the trader.

Hedging and risk transfer represent the mechanism by which risk is managed within the intermediation process. The general principle is that where a transaction results in price exposure that cannot be offset by a precisely equal and opposite transaction, it can be carried at reduced overall portfolio risk by the creation of inversely correlated risk positions which result in an overall reduced level of price volatility at a portfolio level. The reduction in risk equates to a reduction in cost that, through the competitive process, should reduce prices in the market for the relevant services provided to investors.

Where the (expected) marginal revenue captured at the point of entry into a transaction exceeds the total costs of initiating, carrying and exiting the associated hedging position, a hedging strategy will tend to earn a positive return. Costs include:

- bid/ask spreads involved in executing the hedge concurrently, and unwinding it when no longer required;
- transactions fees and commissions;
- operational and settlement costs;
- the cost of the capital associated with carrying any residual basis risks (see below) resulting from the hedged position being imperfect;
- ‘slippage—the expected value of losses incurred through adverse selection during the maintenance of the hedged position, and imprecision in the formulation of the risk reduction;
- ‘drift’—any additional costs associated with the need to rebalance the hedge position to maintain its effectiveness over time.

Where risk is not entirely eliminated, one form of risk has been exchanged for another form of risk of lesser magnitude, which the trader is prepared to carry as an exposure until afforded an opportunity to exit the residual risk. Such residual risks are referred to as *basis* risks. They reflect the tracking error that is possible between the instrument and its hedge. That tracking error may have structural or stochastic properties depending on the grounds for the price correlation of the hedge instrument with the original position, which may itself be wholly or partially deterministic, of strong or weak causal form, or purely derived from an observed time-series correlation (which may be of strong or weak statistical properties, qualitatively robust or spurious, stable or unstable, or exhibit strong mean reverting or gapping tendencies).

Basis (risk) is a generalised concept referring to the risk that offsetting investments in a hedging strategy will not experience price changes in entirely opposite directions from each other. This imperfect correlation between the two investments creates the potential for excess gains or losses in a hedging strategy, thus adding risk to the position. In general, the objective of a hedging strategy is to exchange a volatile first order source of price volatility for a less volatile second order source of price volatility with stronger structural or stochastic properties in terms of fundamental or mathematical underpinnings, mean reversion tendencies, or capacity to be underwritten by existing offsetting portfolio exposure. The overall reduction of risk at each stage of the hedging process can result in a lower overall cost of any capital necessary to support the intermediation function.

By developing more sophisticated hedging strategies that achieve a higher degree of hedging precision by referencing a larger universe of economic input variable, basis risk can be reduced in aggregate. But to the extent that these input variables cannot themselves be perfectly executed and the associated risks eliminated, this is achieved at the expense of creating a larger number of individual basis risks which cumulatively represent a more precise hedge. The isolation of a primary risk into multiple basis risks by hedging provides a mechanism whereby the residual risks are effectively syndicated and distributed across the multiple markets and specialist participants involved in supplying liquidity to the universe of hedging instruments used.

Since the trader can support the price based on a capacity to hedge with greater precision, the process of price-making involves less instantaneous risk, and different types of risk (basis risk versus outright price risk). For a given cost of capital, the trader can therefore afford to make prices based on a lower-threshold expected absolute rate of return, permitting them to make a more competitive price and supplying liquidity to the market at a lower economic rent. This results in tighter prices for the services provided to investors (eg, immediacy).

Anything that impedes this process—whether by introducing structural obstacles to the effectiveness of hedging, adding cost to the hedging and risk-transfer process, requiring risk positions to be held for a longer period of time at any point with the pricing and/or hedging supply chain, or requiring contributed prices to be ‘exposed to danger’ for a longer period without the capacity to change, cancel or affect concurrent hedging—will have the potential to increase the total amount of risk incurred as part of the market-making process.

The best form of hedge is always the elimination of the underlying risk, but this is rarely achievable at any single point in time, so any market-maker is faced with managing a complex portfolio of offsetting risks so as to attempt to minimise the total risk incurred at a portfolio level. This is done by rapidly repackaging primary risk into less volatile basis risks, and fine-tuning pricing so as to exit those basis risks at levels as good or better than those implied by the trader’s original pricing models and slippage/drift assumptions.

Since many basis risks have weak fundamental underpinnings, their use may only be appropriate for very short periods. Others may be structurally very strong, and can be relied on for longer. Different asset classes exhibit a tendency to depend on different types of basis relationship: fixed-income markets and some commodity markets tend to rely heavily on strong structural correlations that can be carried and managed for long periods (even indefinitely) with relative precision; whereas equity and foreign exchange markets tend to rely on loose empirical relationships that are vulnerable to instantaneous change based on external events.

Basis risk can therefore be extremely different at a qualitative level, ranging from robust closed-form arbitrage conditions that can be reliably managed to maturity at predictable cost, to very loose relationships which, while useful over short-term time horizons, have little fundamental underpinning and are vulnerable to instantaneous change or collapse. Accumulation of any form of basis risk on a large scale has the potential to incur material losses should the basis relationship change or collapse; at an industry-wide level, systemic accumulation and collapse of basis risk based on spurious assumptions has often been the cause of major market dislocations.

HFT strategies tend to avoid the accumulation of either primary risk or any material basis risk, as compared with more traditional forms of intermediation, which tended to rely on managing and financing significant basis risk.

In the next section the trading sequences primarily associated with the transactions providing services to investors are described. However, the transactions relating to hedging and risk reduction have the same general features. In order to achieve a reduction in risk, transactions must actually be carried out, and these transactions can take place either by the trader executing against an existing resting order (which is likely to incur higher transaction costs) or by the trader allowing their existing resting order to execute as it gets to the head of the queue in the tick at the touch. The same general issues relating to the positioning, repositioning and cancelling of unwanted resting orders arise.

3 Generic trading sequences used by high-frequency traders

3.1 Introduction

As a result of discussions with the relevant stakeholders, Oxera has established a number of generic trading sequences that appear to be the principle means by which HF trading strategies are implemented (ie strategies that use one or both of very fast trading (low latency) or ‘being fastest’ to create a profit from trading (rather than holding) securities).

Five generic sequences have been identified to cover most of the activities undertaken by high-frequency traders (see Annex 1 for illustrations of all the sequences):

- pure arbitrage—trading the same security in different venues at different prices;

- linked arbitrage—trading securities that exhibit a definitive price relationship(s) that are currently out of normal alignment and which can be expected to revert to their normal relationship quickly (for example, a security and its derivative);

- directional trading—a trading sequence that is based on the (short-term) predicted movement of the price of a single security;

- statistical arbitrage—trading securities that exhibit (or at least have exhibited) price correlations even if there is only a weak, or non-existent, direct relationship between the value of those securities;

- market-making in a single security—directly providing immediacy for investors and (quickly) closing out any positions taken by providing an immediacy service to an investor who would like to undertake the opposite transaction.

In addition to these trading sequences there is another activity carried out by high-frequency traders (and others) that does result in some trading, but the direct objective of the specific order flow to the trading venue is not really trading:

- information discovery—not really a trading sequence, but the practice of sending an order to a trading venue to try to discover if there is demand for that security (either buy or sell) that is not otherwise visible. This activity is often described as 'pinging'.

This is not a definitive list, but, after consultation with the industry, it appears to cover most of the trading sequences currently used at this generic level of description. The rationale for these trading sequences has the potential to explain much of the observed behaviour of the markets in the presence of HFT. In particular, the high cancellation rates of orders at, or close to, the touch are consistent with the rationale for the directional, market making and at least some of the arbitrage strategies, and are consistent with high-frequency traders using short-term price change predictions in their trading strategies.[39] [40]

In the descriptions of generic sequences set out in this section the following definitions, in line with existing industry definitions, have been used.

- *Market order*—an order sent to a trading venue which has the characteristics: buy (sell) a specified volume of a security at the best price available in that venue. This type of order cannot become a passive order.

- *Limit order*—buy (sell) a specific volume of a security at a specific price, or better. This type of order can be either an aggressive order, if it executes against existing passive orders on arrival at the trading venue, or a passive order, if it fails to execute on arrival because there are no suitable existing passive orders at the specified price (or better). A single order can end up being partially aggressive and partially passive if on arrival there are only sufficient resting to orders at the right price to provide a counterparty to some, but not all, of the order.

- *Passive order*—an order that is available in a venue that will not execute against another order that is also currently available in that venue. All passive orders are limit orders, but not all limit orders are passive orders.

- *Resting order*—an order that does not execute on arrival and is not immediately cancelled will end up resting in the trading venue. The order book on a trading venue is made up of the totality of all resting orders. Passive and resting can often be used interchangeably, with the difference between them being that a resting order indicates a state of an order, while passive is more often used to indicate the flow of orders.

- *Aggressive order*—an order that, upon arrival at a trading venue, can immediately execute (all, or in part) against an existing passive order. Market orders are always aggressive, whereas a limit order is aggressive if it executes against an order that has arrived at the venue at the same time, or earlier.

[39] However, in undertaking this research it has not been possible to link the rapid placing and cancelling of orders in ticks a long way from the touch with profitable trading sequences. The issue here is that these orders do not normally result in any executions and, therefore, cannot be part of a trading sequence. To the extent that these orders and cancellations have some other purpose (for example, to confuse other traders or to slow down the trading venue) this is something that has not been captured in this analysis. In addition, there is a limited empirical base for understanding this type of order/cancellation sequence in Europe. In the USA, the particular characteristic of Regulation New Market System (Reg NMS) may mean that this type of observed behaviour there would have a very different impact in Europe where the Reg NMS-type system does not operate.

[40] In addition, Oxera has not analysed any of the rationale for, nor the analytical techniques used by, high-frequency traders to make the price movement predictions. This is beyond the scope of this research.

- *Fill-or-kill order*—a limit order that either executes in full as an aggressive order, or is immediately cancelled. This type of limit order can never be a passive order.

- *Immediate or cancel*—similar to a fill-or-kill order, but this order can partially execute with the remainder of the order cancelled if there is insufficient resting volume (at or better than the specified price) to execute all of it. This type of order can never become a passive order.

Although there are other order types (particularly in relation to orders that can be partially filled) these have not been used in the generic descriptions below. In reality, such refinements would be part of the actual trading sequences used, but these refinements are unlikely to alter the underlying pattern of orders.

In addition to the definitions in relation to orders, the following definitions of market conditions have also been used.

- *Tick*—the price point at which resting orders can rest. Trading venues will generally pre-determine the potential price points that are available (for example, £10.00, £10.05, £10.10). Prices between the ticks cannot be used (for example, £10.03 per security is not a legitimate price). The size of the tick is the gap between one price point and the next. In the example above, this is £0.05 or 5p per share.

- *Spread*—the resting offers to buy and sell will not overlap (if they did a trade could take place). The gap between the best price on offer to buy a security and the best price on offer to sell a security defines the spread. For example, if there are resting offers to buy at £9.95, £10.00 and £10.05, and offers to sell at £10.15, £10.20 and £10.25, the spread is between £10.05 and £10.15, so is £0.10. The spread may also be expressed in basis points, which, in this example, is £0.10 over the midpoint price of £10.10, which is approximately 100bp.

- *Touch*—the ticks at each side of the spread determine the touch. In the example above, these are the ticks £10.05 to buy and £10.15 to sell.

- *Leg (in relation to trading sequences)*—the trading sequences described are designed to bring the high-frequency trader back to their original position of not holding a security. In order to distinguish the sequence of orders designed to achieve this (for example, a *buy* then a *sell*) each order is described as a leg of a trade, usually in the form of first leg, second leg, etc to indicate the sequence.

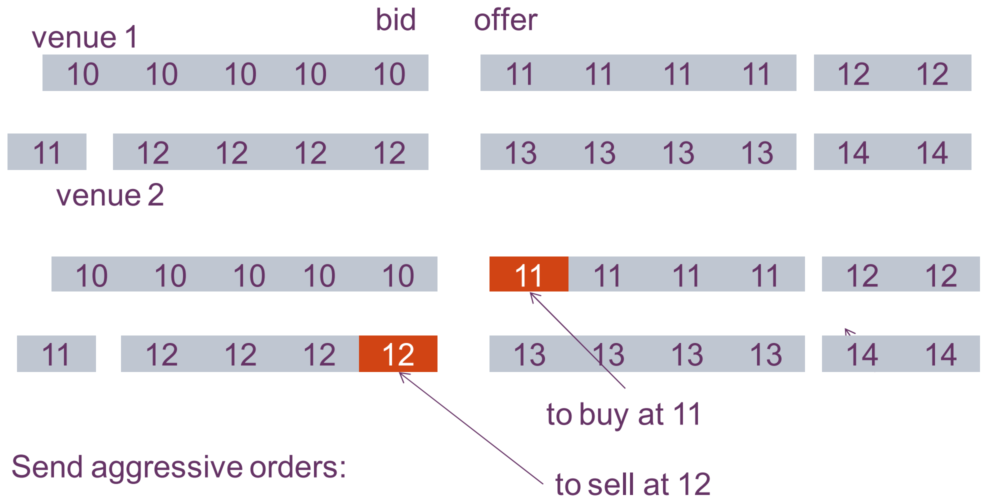
The following sub-sections set out the generic trading sequences, starting with the simplest and working up to the more complex sequences. One aspect of HFT that underlies many of the issues related to this activity is the level of orders and cancellations sent to a trading venue compared to the number of orders that actually execute. In the descriptions below, the conditions under which orders are cancelled is also described as part of the sequence.

3.2 Pure arbitrage—same security traded on different venues

The trader monitors the price of the same security on different venues and when pricing anomalies emerge, the trader attempts to buy the security in the low priced venue and sell the security in the high priced venue.

In its simplest form, this strategy involves the use of aggressive orders that immediately execute with a resting order in each venue (see Figure 3.1 below). In order for this strategy to be successful, the price difference between the venues has to be generally bigger than the spread. If, prior to the anomaly occurring, the prices in the two venues were the same, this would involve the price moving in one venue through multiple ticks before the price adjusts in the second venue.

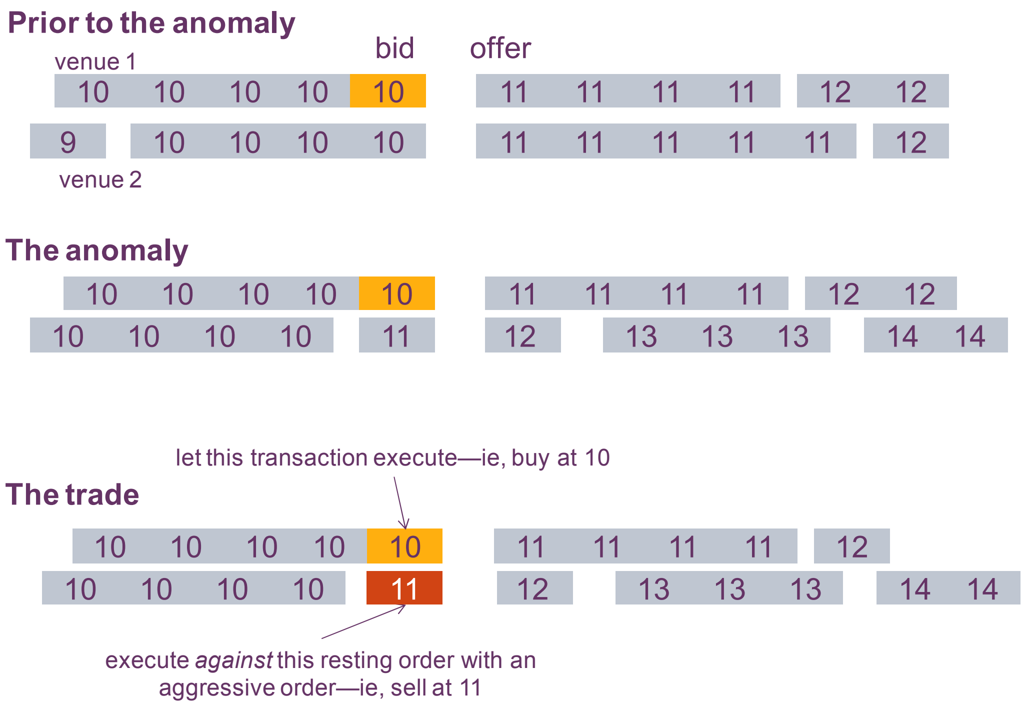
Figure 3.1 Simple arbitrage—aggressive orders



Note: See Annex 1 for how to read these figures.
Source: Oxera.

There are variants of this approach which can exploit resting orders in one or both of the transactions. This involves placing orders into ticks that are not necessarily at the touch in advance of the anomaly arising, and allowing the relevant order(s) to execute (see the section on market making below for a more complete description of this process). Figure 3.2 shows such a sequence, using one aggressive and one passive order.

Figure 3.2 Simple arbitrage using one aggressive and one passive order



Source: Oxera.

Where exclusively aggressive orders are used by the arbitrage trader they are likely to be of the form: 'execute immediately, or if it does not execute immediately then cancel (or fill or kill)' or constructed in a similar way (for example, a cancel sent very quickly—micro-seconds—after the original limit order is sent). Market orders are unlikely to be used as the primary means of execution of this strategy, but if *one* leg of the limit order fails, leaving the trader with a net position, a market order may be used to immediately unwind that net position.

Given that each arbitrage opportunity has a certain maximum value, those traders who are not consistently the fastest are unlikely to be able to use the arbitrage strategy.

Cancellations arise from the failure to be able to execute the required transactions because someone else has got there first.

Resting orders can be used to take advantage of relatively small price anomalies. The trader then has to have one or more passive orders at the head of the relevant queue(s) at the time the anomaly arises. Under these circumstances, the resting order(s) that gets to the front of the queue when the anomaly has not arisen will be likely to be cancelled (just) before it reaches the front. Hence trading sequences using passive orders will generate more cancellations as it is unlikely that each time the relevant resting order reaches the front the right pricing anomaly will be present (indeed, it is unlikely).

Where different venues exhibit a predictable leader/follower relationship, the pricing anomaly may be predictable over short time horizons. In this case, the trading strategy can more easily use passive orders that rest on the relevant exchange. For example, if the price of a security is predicted to rise on the leader venue, a resting order one tick away from the touch can be placed in anticipation. When the price moves into that tick on the leader venue, then before the price has changed on the follower venue an aggressive order can be submitted and, if that executes, the resting order in the primary venue is left in place until that executes. If the aggressive order in the follower venue does not execute (for example, the order arrives 'too late') the resting order in the primary venue is cancelled.

3.3 Linked arbitrage between related securities

Where two securities are linked in a mechanistic or probabilistic way (for example, the underlying security and a futures derivative, or stocks in the same industry, or a constituent stock of an index and the index), price changes in one are likely to signal (or may actually cause) price changes in the other.

As a result, the same price leader/follower pattern may emerge, and price changes in one security can be used to predict price changes in the other. If a trader can react faster than the target counterparty in the market for the follower security to changes in the leader price then there will be a small window of opportunity to exploit that difference. This can be either in a direction—the price of the target security will rise or fall and instigate a buy/sell sequence for that security—or in a form where the prediction is that the prices will come back to their 'normal' relationship, so a sequence of buying then selling, or vice versa, in both securities is undertaken.

The same types of linkages will exist in stronger and weaker forms between single securities and market, or sub-market, indices, and between different single securities. There is, therefore, at least in theory, scope for profitable trading by offering prices based on these linkages. Box 3.1 sets out the type of considerations that are likely to be built into the price-setting strategy of high-frequency traders using this approach.

Box 3.1 Pricing based on index and sector correlation

With any single stock, the fundamental value of the company (the underlying investment in the security) and the way that value is priced in the marketplace at any particular time has a number of correlation factors. Take, for example, Nestlé, a stock in the food sector. The food sector might have a sector trajectory that is independent of the market overall and Nestlé will have a beta[41] relative to the market index, but it will also have a beta relative to other food companies.

When a trader makes the price of Nestlé, he does not make that price solely with reference to Nestlé, he makes it with reference to the general level of the market index(es) of which it is a component, the sector of which it is a component, and the improvement or deterioration of its performance relative to its immediate competitors. The trader could: break the quoting process down to a level of beta correlation to the index; isolate the level of beta correlation between the industry sector and the general index; isolate the level of correlation between Nestlé and the food sector; look at all of the pairs relationships between Nestlé and other stocks in the food sector (eg, Kraft).

The 'pricing recipe' consists therefore of a beta relationship to the index (or even index future), a beta relationship to a food sector index, and a set of pairs

[41] A measure of the volatility, or systematic risk, of a security or a portfolio in comparison to the market as a whole. One can think of beta as the tendency of a security's returns to respond to swings in the market. A beta of 1 indicates that the security's price will move with the market. A beta of less than 1 means that the security will be less volatile than the market. A beta greater than 1 indicates that the security's price will be more volatile than the market.

relationships to Nestlé's immediate competitors.

If, in the course of a trading session, the equity index has not changed at all but the food sector has fallen by 20%, the trader is not going to be making the same price for Nestlé relative to the index, because the fall in the food sector is taken into consideration. And if Nestlé has a short-term momentum that is increasing relative to the food sector, the trader is likely to bid the stock up relative to the rest of the food sector.

In addition to the mechanism the trader uses to work out a theoretical price for the stock predicated on all of the above considerations, externally observable and actionable data allows the trader to—instead of simply taking the outright risk on Nestlé—reduce the risk to a hedge in the index and a hedge in the relationship between the food sector and the index. If the trader was long in Nestlé as a market-maker but already had a short position in Kraft (for example), that strategy may be risk-reducing and not risk-increasing, thereby allowing him to make a more competitive price on the basis that it contributes to risk reduction and need not command a premium to address the associated basis risk.

Variants on this general strategy can be used to exploit pricing anomalies where a security trades in one form in one market (eg, as a normal equity in Europe, priced in euro) and in a different form somewhere else (eg, as an ADR in the USA, priced in US$). A similar opportunity exists in relation to linked foreign exchange transactions where a combination of linked exchange rates leads to trading profit (eg, £ to euro, euro to US$, US$ back to £). As the pricing of each of these securities, or the individual exchange rates, will tend to be moving with only a small degree of independence, any opportunities for profitable trading of this sort will tend to be (very) short-lived. Fast reaction times, and being in the right place in the queue (in some trading venues), will be important in being able to successfully exploit these pricing anomalies.

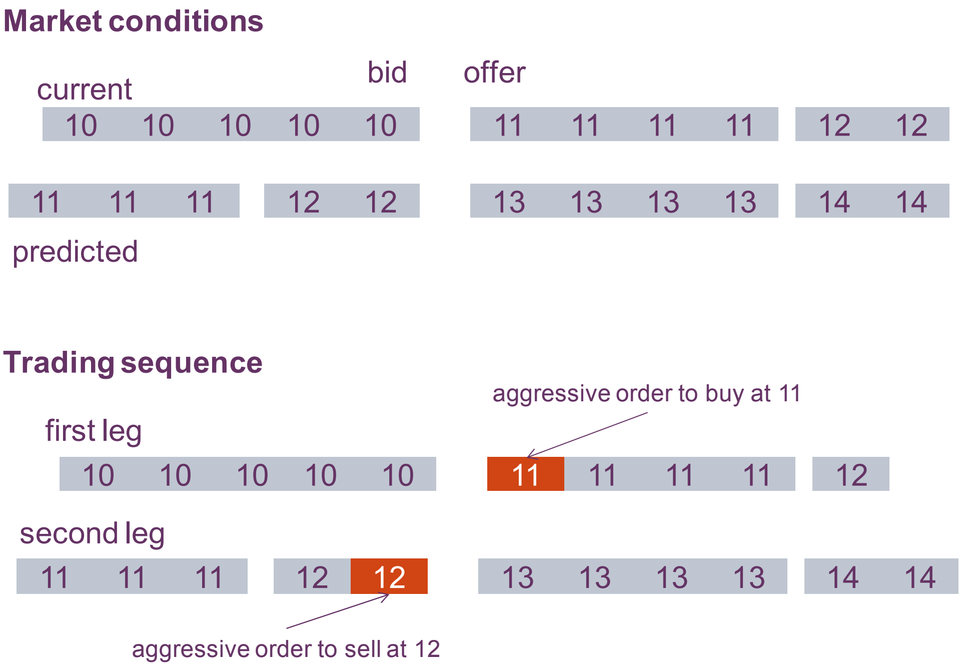
3.4 Pure price movement prediction (directional strategy), on the same venue

The algorithms used by high-frequency traders can make probabilistic, short-term predictions with respect to price movements of securities. Profitably trading these predictions requires that, over the time period of the movement, a complete round trip (eg, buy and then sell) is executed (at appropriate prices). These strategies can be implemented using two aggressive orders, one each of aggressive and passive, or two passive orders. The use of the aggressive orders requires a larger price movement for the trading sequence to be profitable, and the use of pure passive orders takes on some of the characteristics of a 'market making' strategy.

3.4.1 Using aggressive orders only (for example, with a predicted price rise)

In the simple arbitrage sequence using aggressive orders, the price anomaly was sufficiently large to absorb the spread between the resting bid and offers. Similarly, if the (predicted) price change is sufficiently large (and actually occurs) two aggressive orders can be used in this directional trading sequence. The first leg of the round-trip trade consists of sending an aggressive order to the venue that is expected to execute against a resting order that is thought to exist (ie, at the time the order is sent it appears to the trader that there is such a resting order at that venue). This order is likely to be of the fill-or-kill variety. If the price moves (up) a minimum of two ticks plus the spread (ie, if the spread is one tick, the price movement must be at least three ticks and if the spread is two ticks, the price movement must be at least four ticks) then an aggressive (sell) order can be executed at one tick (or more) higher than the price of the initial transaction. Figure 3.3 illustrates the sequence based on a spread of one tick.

Figure 3.3 Directional strategy using aggressive orders only



Source: Oxera.

If the first order fails (ie, by the time the fill-or-kill order arrives at the venue the resting order against which it was designed to execute has either been cancelled or has already executed against another order that arrived ‘earlier’), then no other orders are sent to the venue. The order that has failed to execute can be cancelled (although it has little chance of executing in the short term because there are better prices now visible in the market for any potential counterparty).

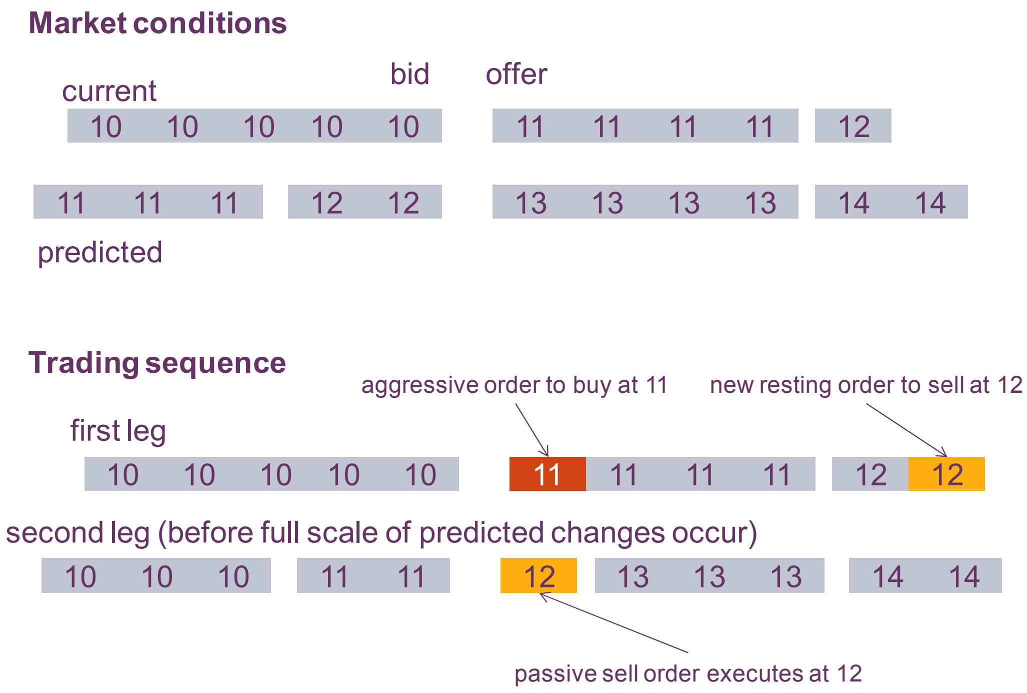
If the first leg does successfully execute, then a second aggressive order is sent to the venue when the price reaches the required tick and the prediction is that such an aggressive order will execute. If that order fails, then a replacement order (at a worse price) may be sent immediately, so as to restore the trader to their original zero holding.

As a result, where aggressive orders are used, the cancellation rate is likely to be driven by how successful the trader is in getting the first leg of the sequence executed (ie, being ‘faster’ than others).

3.4.2 Using a mixture of aggressive and passive orders

The same strategy can be executed using an aggressive order for the first leg and a passive order for the second leg. Given the price/time priority of most venues, at the time the first leg is attempted, one or more resting orders can be placed at the back of the queue, on the other side of the limit order book, in one or more ticks where it is predicted that the price will change to (or through). If the first order does not execute, then the second leg order(s) is cancelled. The simplest version of this sequence is shown in Figure 3.4.

Figure 3.4 Directional sequence using aggressive and passive orders



Source: Oxera.

If multiple orders are placed in multiple ticks then within the timeframe of the movement of the price through the resting orders a decision will need to be taken to cancel any order(s) that are now at a worse price than that which the algorithm is now predicting can be achieved. These orders will need to be cancelled very quickly. In addition, if the prediction of the best price achievable changes (for the worse) one (or more) of these orders may need to be reinstated again very quickly.

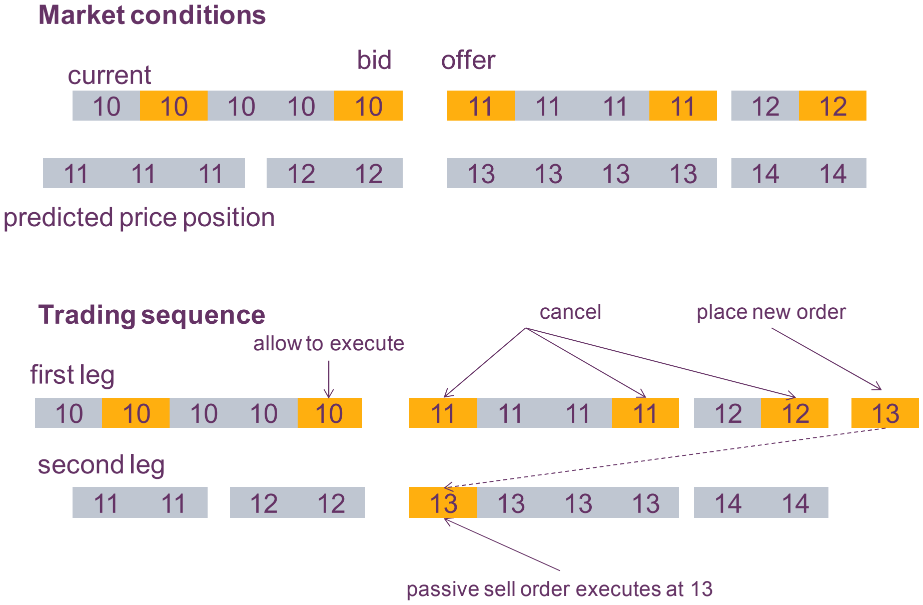
The combination of multiple orders of which only one will, in the end, be needed in a successful trading sequence, and the probability that the initial trade will fail (for example, because that order arrives 'too late'), means that the cancellation rate can be quite high. Even successful sequences may result in multiple cancellations (ie, those passive orders that turn out to be 'in the wrong tick').

3.4.3 Using only resting orders

It is also possible to initiate this sequence using a resting order, if that resting order is allowed to execute only when the algorithm indicates that the next price movement(s) is in the appropriate direction. Using this approach a trader would maintain resting orders in multiple ticks on both sides of the spread. Those at the touch would only be allowed to execute when conditions were 'right' and, as a result, orders at the touch would be cancelled and immediately replaced if there was a danger that they would be at the head of the queue and execute at the 'wrong' time. In addition, where short-term price variations (measured in ticks) are significant, resting orders outside the touch may also be vulnerable (in a probabilistic way) to execute at the 'wrong' time. Traders may, therefore, have reasons to cancel and immediately replace orders beyond the touch as well, so as to send them to the back of the queue within that tick.

This approach to the strategy can, therefore, result in large numbers of cancelled orders where there is a danger of executing at the 'wrong' time. Traders are positioning themselves to execute a sequence of trades at the 'right' time, but in order to be optimally placed when the opportunity arises, resting orders (especially at or close to the touch) may require fairly constant repositioning. Figure 3.5 shows a sequence where the resting orders are being maintained on both sides of the spread.

Figure 3.5 Directional sequence using passive orders only



Source: Oxera.

The round trip in this example involves allowing an existing order to execute, the cancellation of three existing orders on the other side of the spread and the placing of a new order (in this case at a higher price), which is then allowed to execute. In addition, in the face of a predicted price change of this sort, and the failure of the first leg to execute, the three existing orders on the other side of the spread would have to be cancelled anyway. With a number of high-frequency traders operating in this mode, and a limited number of (probably investor) counterparties that want to trade on the aggressive side of the transaction, the ratio of executed orders to cancelled orders can be quite high.

The strategy of placing a number of orders throughout the order book (frequently described as 'layering') is often done so as to establish positions so that the trader can obtain, when needed, higher time/price priority in the order book. However, this inevitably exposes the trader to greater risk that under certain conditions, the sudden arrival of a large aggressive order will cause many of the resting orders to execute at the same time, and before (even for the fastest trader) these orders can be cancelled.

Layering can be an entirely legitimate process designed to achieve a better average price by being at the right place in the tick at the right time. However, it can also be used to try to create a false impression of the order book and 'paint a picture' of deeper liquidity than is really intended to be made available. The potential for the order book to be 'swept' by an aggressive opposing order always exists, however, so this tactic still contributes liquidity, albeit of lower quality than a genuine intention to execute at both prices.

When the incoming aggressive orders that are (rapidly) changing the tick in which orders are executing is not a single order, then the speed of reaction time to cancel the resting order(s) will have a significant impact on the risks of a resting order executing at the wrong time. As a result, successfully maintaining these multiple positions in multiple ticks and trading on predicted (short-term) price movements is effectively restricted to those with very fast reactions and very fast connections to the trading venues. In practice, therefore, this approach is not really available to anyone without competitive infrastructure.

When (legitimately) layered orders are themselves conditional on external pricing inputs (ie, not just the information on that trading venue relating to that specific security) or other risk criteria (eg, the changes in the traders' position in other securities), a high propensity to cancel is also likely to be exhibited, as the fundamental motivation for the placement of the order can change as a result of input from many sources, and these inputs may themselves be changing frequently. Prohibiting, or restricting, layering in this form may result in a material reduction in total liquidity available, and could also impact the function of the market as a means of information transfer.

3.5 Statistical arbitrage

Similar to linked arbitrage, statistical arbitrage involves more complex sequences across more securities. However, the same choice of trading methods is available. Sequences can be executed as aggressive orders and, where available, resting orders. Hence, where resting orders are used, repositioning may be required to enable the trader to be optimally placed when the 'right' time arises (as predicted by the algorithm or as observed by the current pattern of prices). Given that multiple securities are likely to be involved in any execution sequence, where resting orders are part of the trading sequence, the volume of repositioning orders (ie, cancellations) can be very high.

3.6 Market making in a single security (using HFT techniques)

The trading sequences described above can involve significant, or even exclusive, use of resting orders. These may have the effect of providing liquidity, and as a result, these trading sequences can be characterised as high-frequency traders being market-makers. However, there is a more pure form of more traditional market making that some high-frequency traders adopt. In this strategy, the trading sequence is designed to provide a round trip that crosses the spread, carried out over very short time periods.

In the securities that are most used by high-frequency traders the spreads tend to be very narrow. A small adverse price movement within any round-trip sequence will therefore tend to wipe out any profit from that round trip where there is a narrow spread.

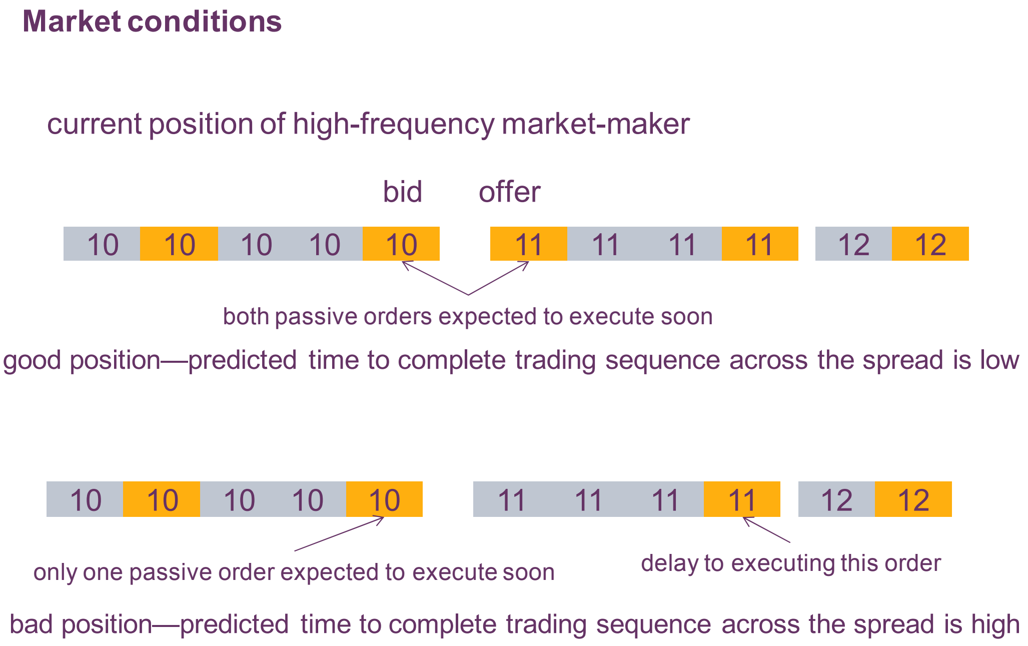
Consequently, market making across this narrow spread requires that when a sequence of trades is initiated, the probability of an adverse price movement is low and, better, the probability of an advantageous price movement is high. Under these circumstances the desirability of initiating a sequence of trades starting from a resting order is likely to vary significantly over very short time periods. This can be directly as a result of the algorithm indicating that over the next (short) time period the price movement from any particular starting point would be adverse, or that the forecast volatility of the next short time period is such that starting from either side of the spread would be high-risk.

The result is that undertaking this type of market making involves significant and frequent repositioning of orders, particularly at the touch, as the conditions for initiating a trading sequence across the spread change rapidly. If the short-run prediction is, for example, for a price to fall, an existing resting order to buy at the touch would be cancelled. (The existing resting order to sell would not be cancelled, because the short-term movement would be favourable to a sequence of trades starting with this trade.)

Predictions of higher volatility could result in resting orders on both sides of the spread being cancelled until the short-term predicted volatility reduces. While the higher volatility is predicted the spread may widen (for example, as a result of the cancellations above) and at this wider spread the market-making sequences described above may return to being profitable. As a result, changes in the predicted short-term volatility of the price of a security can result in quite significant cancellations of existing orders at the touch (as volatility is predicted to increase) and repositioning orders inside the (then) touch (as volatility is predicted to fall).

In addition, because the complete trading sequence involves two transactions, even if a favourable movement is predicted it may be risky to let the first leg take place without the resting order for the second leg being in a favourable position in the time queue in its tick. Figure 3.6 below shows two patterns of resting positions where, although no price movement is predicted, a high-frequency trader market making may cancel, or go ahead with, an attempted market-making sequence.

Figure 3.6 'Good' and 'bad' positions for market making


Market conditions
current position of high-frequency market-maker
bid
offer
10 10 10 10 10
11 11 11 11
12 12
both passive orders expected to execute soon
good position—predicted time to complete trading sequence across the spread is low
10 10 10 10 10
11 11 11 11
12 12
only one passive order expected to execute soon
delay to executing this order
bad position—predicted time to complete trading sequence across the spread is high

Source: Oxera.

Complex patterns of repositioning of orders, especially at and close to the touch, are likely to result in (very) high cancellation rates relative to execution.

3.7 Combinations of trading strategies

In reality these trading strategies are not necessarily exclusive. Indeed, successful strategies are likely to combine all of these ways of exploiting either short-term pricing anomalies or short-term predictions of price movements or volatility changes. Additionally, all of the above trading sequences are likely to be combined with the use of orders that are specifically designed to generate (additional) information about the state of the market/order book (see section 3.8). The different types of information are then combined and used to drive the decisions on trading sequences.

An example of the types of decision being made in relation to market-making when more complex strategies are being pursued is set out below in Box 3.2.

Box 3.2 Complex market-making

Traders who make two-way prices in equities markets never earn the whole spread because they expose themselves to the risk of being traded against by better-informed traders all the time. Over time, if the market-maker can on average capture a fraction of the spread that exceeds his total cost of slippage, he will make money.

The way slippage works however is that one tends to earn small amounts regularly but lose large amounts irregularly. Although the market-maker is picking up a spread with very high frequency, there is always the possibility of some exogenous event, which could not be interpreted from available data, occurring and wiping out those profits and more. However, if the market-maker is doing a good job, over time he will make more than he will lose.

HFT firms tend to have that sort of profile, which is why they pull out of the market in times of stress, when they do not know what is happening and they do not wish to expose themselves to loss.

When market-making was done with relatively low time precision[42] by big financial institutions, they often fulfilled a dual role: providing transactional liquidity and an underwriting capacity to the market (ie, by continuously quoting, even during times of stress). Although these two roles were poorly abstracted, the old-style market-makers were able to trade profitably in this dual role, largely because they had an information advantage, and the associated spreads where wide, meaning the cumulative bid/ask capture represented a large pool of revenue against which to offset periodic losses.

Today, HFT firms do not generally have any intrinsic informational advantage (such as an 'inside picture' on customer order flow or investor sentiment) other than that conferred exclusively by fast access to the markets. Generally their role does not involve forming a directional opinion about where the stock price is going, but rather to facilitate the three types of intermediation: time, price and/or form. However, a successful market-making strategy may have a component that is trying to achieve a change in a net position in a security based on some other business objective, such as an underlying objective to accumulate or reduce a particular risk position, which causes the market-maker's price (or willingness to trade on a particular side of the spread) to be 'skewed' to make the bid or ask side somewhat more competitive, or to differentially allow resting orders on different

[42] Prices were changed relatively infrequently.

sides of the touch to actually execute. This directional overlay may be in respect of the primary instrument, or any of the underlying basis relationships from which its price was derived, such that the price made by that market-maker has a tendency to be particularly competitive on one side of the spread. Market-makers with offsetting underlying objectives may incur sufficient skew in the resultant prices that their prices match with one another (and hence they trade with each other), with both prices being consistent with the parties' objectives.

This helps explain the high volume of transactions seen between wholesale market participants that are independent of end-user investor involvement, and also demonstrates why transactions are not a 'zero-sum game' (at least in terms of the reduction in risks and therefore costs facing the market-makers). With very tight bid/ask spreads, the costs associated with the transfer of offsetting risks can be very low, and so the volume of such risk transfers can grow rapidly, since the threshold transaction cost incurred in offsetting a risk is reduced, allowing risk to be reduced or eliminated.

The market-maker may, for example, have an initial 'pairs' position (a form of basis position) whereby, because he is making markets in multiple stocks, he might be longer in Intel and shorter in AMD than he wants to be, in which case he will tweak one of the secondary parameters, to trim out the bias in the way positions are being accumulated, or to deliberately bias the inventory accumulation in order to achieve a directional objective.

There are perfectly valid reasons to have a bias in how one makes markets, but the subtleties are typically expressed in those second- and third-order pricing components, which are themselves not volatile. The successful market-maker will generally have an algorithm that is basically a formula plus alpha[43] plus beta. He is trying to rely on the more stable pricing parameters, which tend to be highly mean-reverting, in order to manufacture liquidity in the primary instrument being quoted.

3.8 Information discovery (pinging)

The trading sequences described above relate to the execution phase of HFT. Information is gathered about the condition of the market and (generally) predictions are made about the likely condition in the next (brief) time period. Trading sequences are then undertaken, or not, according to the analysis.

In addition to this (relatively) passive analysis of market data, those wishing to undertake trades may also actively engage with the market in order to (primarily) generate information about the market conditions rather than actually undertake the trade. With respect to HFT, a significant example of this is the practice of sending orders to the trading venue to try to establish if there are existing counterparties to a trade at a particular price that are not visible. That information (ie, that there exists, or does not exist, a matching counterparty to the order sent) is then used as part of the information underpinning the decision to proceed (or not) with a trading sequence.

[43] A measure of performance on a risk-adjusted basis. Alpha takes the volatility (price risk) of a portfolio or fund and compares its risk-adjusted performance to a benchmark index. The excess return of the fund relative to the return of the benchmark index is a fund's alpha.

For HFT, the primary manifestation of this approach is the practice of 'pinging' orders. In a trading lit venue with the potential for hidden orders (ie, orders that are on the order book, but are not visible) a ping order is sent at a price inside the spread. If it executes, then that indicates the presence of resting hidden orders at that price. If it fails to execute, then there are no hidden orders at that price (or better).[44] Because the objective of the ping is to try to discover the presence of hidden resting orders, there is no guarantee that the *execution* of the order is part of the trading sequence that would be instigated having made that discovery. Indeed, the opposite is more likely—ie, that the presence of the hidden order would instigate a trading sequence that started with a trade in the opposite direction to that of the ping order itself. As a result, the failure of a ping order to execute against a hidden order is likely to mean that the trader does not want that order to execute against any new incoming orders. However, because the ping order is within the spread there is a very high probability that such an order that fails to execute on arrival would execute very quickly as it will represent a 'better' price for counterparties compared with the resting orders that existed just prior to the ping order arriving. As a result, these pinging orders (when used in this way) will tend to be cancelled very quickly, if not immediately (for example, sent as an immediate or cancel or fill-or-kill order).

For the high-frequency trader these pinging orders are valuable when:

- the information gathered when they do execute is greater than the cost to the trader of them executing, combined with;

- the information gathered when they do not execute is greater than the cost of them executing by mistake, not against a hidden order but against an incoming marketable order arriving after the ping order has arrived.

High-frequency traders are not the only users of pinging orders. A trader wishing to change their net position may wish to (try to) understand the disposition of available resting orders, including hidden orders within the spread, in order to optimise their order routing and to minimise market impact. However, in this case a ping order that executes is more likely to be in the same direction as the wanted change of position. As a result, for these traders, a ping order that fails to find hidden resting orders, but which then executes against the next incoming marketable order, would not necessarily be seen as a 'wasted' order.

4 Market-making requirement

4.1 The proposed rule

The MiFID consultation paper proposed that:

> Market operators would be required to ensure that if a high frequency trader executes significant numbers of trades in financial instruments on the market then it would continue providing liquidity in that financial instrument on an ongoing basis subject to similar conditions that apply to market makers;[45]

[44] This over-simplifies what actually happens, as there are hidden order types which will only execute if the totality of the order can execute. Under these circumstances a small ping order may not execute as it would only take up part of the hidden order. A non-executing ping order would then not necessarily indicate that there was no hidden counterparty at that price.

[45] European Commission (2010), 'Review of MiFID', December 8th, p. 16.

In the Commission's final version of its proposals, this has been reworded as follows:

> An algorithmic trading strategy shall be in continuous operation during the trading hours of the trading venue to which it sends orders or through the systems of which it executes transactions. The trading parameters or limits of an algorithmic trading strategy shall ensure that the strategy posts firm quotes at competitive prices with the result of providing liquidity on a regular and ongoing basis to these trading venues at all times, regardless of prevailing market conditions.[46]

One of the differences between the original proposal and the final version is that there is no longer a reference to the fact that these high-frequency traders would be subject to 'similar conditions that apply to market makers'.

4.2 The Commission's impact assessment

According to the Commission, the rationale behind the proposed rule is to ensure that high-frequency traders provide meaningful liquidity at all times and would contribute to more orderly and liquid markets and mitigate episodes of high uncertainty and volatility. In other words, the ultimate aim is to prevent market panics and crashes from happening. The Commission summarises the advantages and disadvantages of the proposed rule as follows:

> The advantage of implementing this option would be to ensure that high frequency traders cannot abruptly enter or leave the market for an instrument resulting in a sudden increase or decline in liquidity for that financial instrument. For example if there were adverse market conditions a withdrawal from the market could cause a sudden drain in liquidity which could exacerbate price movements and volatility for an instrument.
>
> A disadvantage could be that high frequency traders may refrain from participating in the markets as they would not want to take on liquidity provision obligations, especially in adverse market conditions.[47]

The way the proposal is now worded, however, means that all AT will be required to use an algorithm that 'posts firm quotes at competitive prices ... at all times'. Although all HFT can be seen as being algorithmic, not all AT is high-frequency. Discussion with market participants suggest that there is very little trading that is purely manual. Agency orders that result in long-term changes in net positions are very often executed using a (non-high-frequency) algorithm. Non-high-frequency market making may also use algorithms to maintain the market-makers' market position, rather than relying on direct human intervention to update those positions as their resting orders execute and additional information is incorporated into the security's price. As a result, the majority of market participants would appear to be caught by this requirement to post firm quotes at competitive prices at all times. The impact of such a requirement would, therefore, extend significantly beyond high-frequency traders. This does not seem to have been taken into account in the Commission's impact assessment.

4.3 How would the rule work in practice?

It is useful to bear in mind that market-making requirements were never introduced with the intention to prevent panics or market crashes, nor to maintain artificially the old price of a security once additional information on that security has become available and it now has a

[46] European Commission (2011), 'Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p. 70.

[47] European Commission (2011), 'Commission staff working paper, Impact Assessement accompanying the document: Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p.128

new equilibrium fair price. They were introduced with the aim of providing immediacy—ie, ensuring that a buyer in the morning does not need to wait until the afternoon to find a seller.

The economics of successful market making require that market-makers are rewarded for the risks they take in providing a counterparty at a time when a natural counterparty is unavailable. As executed prices change as a result of changes in buying and selling pressures, market-makers face a risk that any particular price change represents a change in the fundamental equilibrium price of the security, and if the market-maker continues to transact at their currently offered prices some or all of these transactions will prove impossible to unwind at some later date across a positive spread.

In highly uncertain periods when it is unclear if price movements are random or fundamental, the potential risk facing market-makers increases. This applies to high-frequency traders acting as market-makers, as well as more conventional market-makers. Hence, under conditions of uncertainty, market-makers will widen the spreads they offer, and thus increase the costs of immediacy for any trader wishing to execute a trade. As a result, under conditions of (extreme) uncertainty, all market-makers are likely to withdraw effective liquidity from the market. In general, this is also reflected in trading platforms' rules. Depending on the exact rules in place, at the request of a market-maker, trading platforms may suspend or vary market-maker obligations. For example, they could occasionally allow market-makers to relax spreads when an individual security is subject to wide price movements. Furthermore, market-makers are typically required to make the market only for a certain period of time. For example, the LSE requires market-makers to maintain their quotes for at least 90% of continuous trading during the mandatory period.

In addition, circuit breakers are designed to suspend trading in conditions of very high market uncertainty because of the risks that trading takes place at clearly unrealistic prices. Where the fair price really is unknown, it would seem inconsistent to require algorithmic market makers to continue to trade at 'competitive' prices when, by definition, these are unknown.

In other words, it seems that final version of the Commission's proposal is requiring something from high-frequency (and, it would appear, a large number of other) traders that is not even normally required from conventional market-makers—ie, posting firm quotes on a regular and ongoing basis to trading venues *at all times, regardless of prevailing market conditions*. Making the market under adverse market conditions can involve significant risks and could result in the bankruptcy of the trading firm. High-frequency traders (and any others caught by this requirement) may not be willing to subject themselves to the proposed rule and may therefore decide to withdraw themselves from the market all together. As a result, the market would revert to manual trading, which would be likely to increase the cost to brokers of executing any particular trade (and, indeed, to other participants in the value chain). To reflect these increased unit costs, spreads would be expected to widen. This seems to be unlikely to be the objective of such a rule.

If, more realistically, algorithm-using market-makers are not required to maintain prices in all market conditions and are therefore protected from the situations where very large losses are a significant risk, it is likely that high-frequency market making would be (ie, remain) a profitable business, or at least non-loss making. Under these circumstances the impact of the proposed requirement on the outcome of the market may be (very) limited.

4.4 Potential general impact of the proposed rule

The impact of this proposal will depend critically on how the requirement to ‘continually post firm quotes at competitive prices’ is interpreted. At one extreme, there will be significant additional risk applied to algorithmic trading if the requirement is interpreted to mean that at all times the market is open, every algorithmic trader has to offer to buy and sell a security across a spread that reflects the usual spread for that security. At times of increased uncertainty those required to maintain bid and offers across the usual spread are likely to find that they enter into a significant number of trades where the subsequent track of prices will have moved against them and which they will not be able to unwind without loss.

As explained, this could have the effect of making AT (including buy-side trading using computer-controlled execution) non-viable because of the requirement to provide firm bid and offers at competitive prices being too risky. All trading could, therefore, return to manual trading. If, as seems likely, the return to manual trading increases the total costs of intermediaries (because they have to substitute people for computers) this increase in intermediaries’ costs will be paid for by investors (who will see a lower net return) or companies (who will see an increase in the cost of capital which will then be passed on in the form of higher prices in the end product market).

If, on the other hand, the interpretation of the requirement is designed to mimic the requirements applied now to official market-makers, so that there is still the possibility of market making using computer-driven trading sequences being profitable, then, as explained, the impact on HFT sequences is minimal.

In general, this arises because any HFT sequence algorithm is likely to be able to have a market-making module bolted on to it that will *display* firm bid and offers at competitive prices under normal market conditions, and to widen the spread offered or withdraw from the market under those market conditions where bid and offers across a narrow spread become highly risky (as traditional market-makers are generally currently allowed to do). Such a module could be designed to actually execute very rarely, if at all, by the repositioning of these (resting) orders before they reach the head of the queue. In addition, if the definition of ‘competitive’ price includes prices one or more ticks away from the touch, then ensuring that actual execution is limited, will be easier.

Unlike those intermediaries who wish to make money out of market making, those who simply wish to meet this requirement in order to pursue other strategies will generally only wish to avoid losing money, and therefore will not be concerned to execute trades, at some point, for the purpose of market making.

4.5 Impact on trading sequences

4.5.1 Market-making strategies

The high-frequency market-making sequence already involves posting competitive bid and offers as the sequence is successful only when some execution of trades actually takes place. The impact of the weaker form of the rule (ie, the one which allows widening of spreads and temporary withdrawal of bids and offers under conditions of high uncertainty) will therefore be minimal. As indicated above, the very strong version of the rule may cause all AT to cease.

4.5.2 Directional strategies

Many of the directional strategies already result in resting bids and offers being placed on the touch or in ticks close to the touch. Many of the bids or offers needed to meet the weak version of the rule will therefore already be placed by the high-frequency trader using the market-making trading sequence. However, some additional resting orders may be required to meet the 'at all times' requirement, but these additional bids and offers are unlikely to be wanted to be executed, so are likely to be repositioned whenever there is a risk that they might execute at the wrong time. If there is an impact, therefore, it may be to increase the number of order cancellations relative to executions.

4.5.3 Arbitrage strategies

Simple

Resting orders close to, or at, the touch are part of the trading sequence for arbitrage trading strategies. Hence, in many cases, many of the resting orders that would be needed to meet the market making requirement (ie, to maintain competitively priced bid and offer resting orders in the relevant security) will already be in place. However, as with directional strategies, there may be a requirement to place more resting orders into the market to meet the 'at all times' requirement. In addition, where the trading sequence uses marketable orders (ie, orders that can be executed immediately) it may be necessary to add resting orders to that market, but in such a way that these orders generally do not execute (ie, always repositioned before they get to the head of the queue). With this strategy, therefore, the most likely impact of the weak version of the rule is to increase the number of orders, and increase the number of cancellations relative to executions.

Linked and statistical

In the case of linked and statistical arbitrage strategies, more markets and trading venues are likely to be involved, with more complex sequencing of trades. In general, to meet the proposed rule, market making will be required in each security and each trading venue involved. In the weak version of the rule this will be likely to mean placing bid and offer orders for each venue/security that might be needed, and repositioning them before they can execute (except if they turn out to be the specific transaction wanted). Again, therefore, the likely impact is to increase the number of orders and cancellations relative to executions.

Pinging

If a pinging order is sent to a venue within which one (or more) of the other high HFT sequences is to be undertaken, the market-making requirement will already have been triggered in order to be able to execute that sequence(s). If the pinging order is, however, aimed at a venue within which the high-frequency trader will not be trading (ie, designed to elicit information about the state of the market in that venue, but where the high-frequency trader will not want to actually execute the relevant trading sequence), it is possible that the market-making requirement in the venue to be pinged could be triggered. If it is possible to meet the requirements in these venues, then additional orders (and cancellations) could be expected. If it is not possible to meet these requirements in these venues (for example, the venue operates in such a way that all trades in the venue execute at the midpoint of a primary market, so there are no resting orders on each side of the spread), this may not be possible. As a result, computer-generated pinging orders to that venue may not be possible. However, it should also be noted that in these circumstances computer-generated normal orders may also not be possible, because to use them would also require those intermediaries using computer trading to become market-makers in that venue.

5 Minimum tick size

5.1 Requirement

The MiFID review proposes that implementing measures could further specify minimum tick sizes—it does not prescribe specific minimum tick sizes at this stage. Needless to say, the minimum tick sizes would be applied to all trading and not simply automated trading or HFT.

Tick sizes refer to the minimum increment by which the price of a share (or other security) is allowed to move up or down on an exchange. In the past few years, tick sizes have come down significantly, partly driven by competition between trading platforms in an attempt to attract trading volume and liquidity.[48]

5.2 The Commission's impact assessment

The MiFID review consultation document does not explain the objective of this proposed rule, so it is not possible to analyse directly what market failure this proposed regulation is aimed at, nor is it therefore possible to directly evaluate if an alternative intervention would be available. However, given the form of the rule (minimum tick size) the general objective would seem to be that in the absence of the rule tick sizes would become (or currently are) 'too small'. If the rule is to have an impact it would seem to need to either:

- increase the tick size that would otherwise occur; or (as a side effect)
- coordinate and standardise tick sizes for the same security trading in different venues.

It is possible that the Commission considers that there is a need to impose a minimum tick size in Europe if it considers that tick sizes have become (or could become) too small, and that this small size itself causes direct harm to the capital market. However, given the context of this proposal, it is also possible that the proposed rule is aimed more generally at curtailing HFT, or at least certain trading strategies that are implemented using HFT. As explained below, certain trading strategies implemented using HFT become possible or more attractive when tick sizes are small. Empirical analysis also suggests that securities with smaller tick sizes are associated with higher volumes of HFT.[49]

5.2.1 Harmonising tick sizes in Europe

Costs—if the objective were to set optimal tick sizes in Europe, then there are unlikely to be any significant indirect costs. Regulators and industry would incur some costs in conducting

[48] See, for example, *Financial Times* (2009), 'LSE bows to tick size pressure as war erupts', August 17th.

[49] See, for example, Jarnecic, E. and Snape, M. (2010), 'An analysis of trades by high frequency participants on the London Stock Exchange, Working Paper', June.

research to determine the optimal tick sizes. Although not insignificant, these are likely to be limited.

Risks—determining the optimal level of tick sizes is far from straightforward and there is therefore a risk of setting tick sizes at sub-optimal levels. For example, if tick sizes were to be set too large, not only would it would potentially curtail HFT and the potential benefits of HFT but it could also reduce other benefits that could arise from tick sizes smaller than that prescribed.

Explanation—empirical analysis in the literature has indicated that smaller tick sizes can intensify competition between market-makers and liquidity-providers and therefore attract more liquidity and tighten spreads.[50] However, it should be noted that the empirical studies in the literature measure the impact of a reduction in tick sizes in the pre-HFT era and therefore do not take into account the impact of HFT. In other words, it is possible that the impact of (further) reductions in tick sizes in the current environment of substantial volumes of HFT may be different.

Very small tick sizes may also have negative effects. First, smaller tick sizes mean that the liquidity is distributed across more price points—liquidity at the touch and every other price point may be lower. Although total liquidity (measured by taking into account the liquidity at all price points) may increase when tick sizes are reduced, it has been suggested that with a smaller tick size, the posted liquidity may be more evanescent—ie, liquidity may no longer be there when the order reaches the market.[51] In other words, there is less certainty to traders about the amount of liquidity actually available at any point in time.

Second, smaller tick sizes may have a negative effect on the price–time priority model under which all major trading platforms operate. Under the price–time priority model, orders that are more competitively priced have higher priority of execution, and given two orders that are equally priced, the submitter of the first order is rewarded with a higher priority of execution at that price level. This effect on the price–time priority model was articulated by Winterflood Securities in its response to the MiFID consultation:[52]

> If tick sizes are appropriate, this [the time-priority model] is an orderly approach, and it is also fair because the submitter of the first order is rewarded (with higher time priority) for the informational risk that he assumes, since any other market participant might decide to outbid him. However, any other participant benefiting from the information (by incorporating that information in the pricing model) and deciding to outbid him, will have to pay a premium in the form of a higher price, and the minimum cost of that premium is the tick size. Hence, should the tick size of the venue be excessively small, its participants can decide to outbid anyone at a cost (the tick increment) that may be economically irrelevant. It is, for all practical purposes, a free option on time priority. Because of the above, the Price-Time Priority model is not applied fairly, and the market as a whole becomes unfair.
>
> In addition to the above, the adoption of inappropriate tick sizes might degrade the usefulness of the order book as a price formation mechanism.
>
> Inappropriately small tick sizes will have two effects in terms of order book structure:

[50] See, for example, Chan, K.C. and Chuan-Yang Hwang (1998), ‘The Impact of Tick Size on Market Quality: An Empirical Investigation of the Stock Exchange of Hong Kong’, October.

[51] Cheuvreux (2010), ‘Navigating Liquidity 5’, Global Research, December.

[52] Winterflood Securities (2011), ‘Review of the Markets in Financial Instruments Directive’, February 2nd, p. 5.

- Less liquidity will be posted in order books because, at any given price level, time priority can be bought virtually for free (as explained above)
- The liquidity that is posted in the book will be layered at much thinner levels.

Turquoise summarises the potential negative effects as follows:[53]

> Where the tick size is too low, the cost of setting a new best bid/offer is small, and so large orders are more prone to being "stepped ahead of". This reduces the incentives to display size in the public markets, continuing the trend towards smaller order and trade sizes and more frequent data updates.
>
> Lower liquidity (shorter queues) at each price point, combined with a number of competing order books for each security, might also dilute the incentives to leave orders in the market for a period of time so as to reach the front of queue – and without such an incentive orders will tend to have a shorter duration – once again fuelling faster market data update rates.

In summary, there is likely to be an optimal tick size for each security. However, there is no guarantee that tick sizes will be set 'automatically' at the optimal level by the market. As explained, trading platforms have an incentive to compete on tick sizes and although there have been attempts to harmonise tick sizes in Europe, this has only been partly successful.[54]

Determining the optimal tick size is far from straightforward in itself and is further complicated by the fact that different parties have different interests. It may therefore be more appropriate for a regulator to determine the optimal tick size. This is a challenging task and would require industry consultation.

5.3 General impact of the proposed rule

The interpretation of a requirement of this sort would be that it would impose a minimum tick size that is larger than would otherwise be the case and/or that trading venues would have the same tick size for the same security.

Before analysing the impact of increasing the tick size on trading sequences, it is interesting to look at a number of 'natural experiments' (in the USA) that have occurred whereby the effective tick size of a security in which significant levels of HFT have been occurring, and continues, has changed. A recent example is the reverse stock split of Citi Group that occurred after close of trading on Friday May 6th 2011 (so starts trading in the new form on Monday May 9th), and which was announced on March 21st).[55]

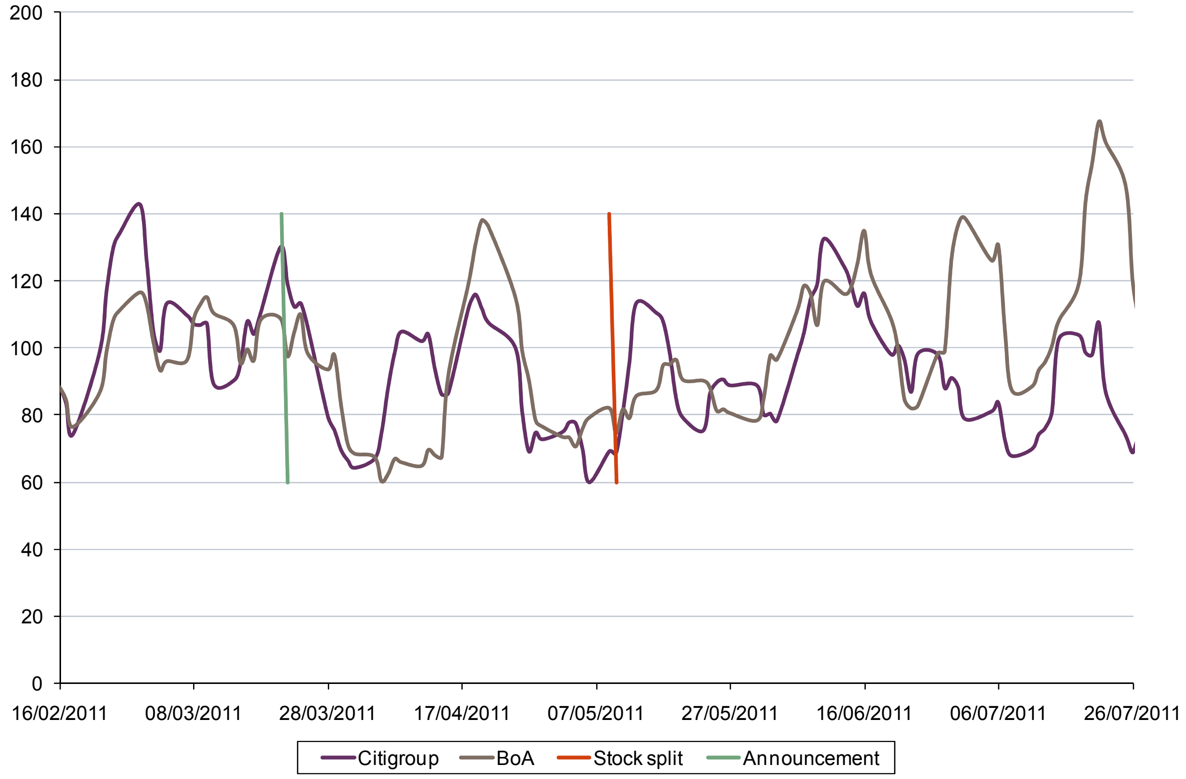
Figure 5.1 below shows the total trading volume of split adjusted shares from the end of January 2011 to the end of July 2011, along with the volume of share trading in Bank of America, both indexed to the average volume of trading throughout that period.

[53] Turquoise (2011), 'Turquoise Talks—Size Matters…', January 27th.

[54] FESE (2011), 'Updated FESE tick size tables as of September 2011', September, available at http://www.fese.be/_lib/files/UPDATED_FESE_TICK_SIZE_TABLES_AS_OF_SEPT_2011.pdf.

[55] See: Citi (2011), 'Citigroup Announces Reverse Stock Split: Intends to Reinstate Common Stock Dividend of $0.01 per Share', press release, March 21st., available at: http://www.citigroup.com/citi/press/2011/110321a.htm.

Figure 5.1 Volume of trading in Citigroup and Bank of America shares over the period of the Citigroup reverse stock split. Normalised to 100 as the average trading volume during the period



Source: Trading volume data taken from Yahoo Finance website. (http://finance.yahoo.com/). Oxera analysis.

Prior to the split, the stock traded in the $4–$5 per share price range, afterwards in the $40 per share price range, declining slowly. Both pre- and post-split the tick size is 1c. Hence the tick size changes from around 20 basis points (1c in $5) to around 2 basis points (1c in $50). This change in the effective tick size is significantly larger than any impact likely to come out of any regulatory intervention in Europe, where the tick sizes of liquid stock tend to lie in the range of 1–10 basis points (depending on the price of the stock).[56]

Notwithstanding the change in tick size of an order of magnitude, the total volume of transactions (measured in terms of value traded) remains remarkably similar. This would suggest that any imposition of a realistic minimum tick size is unlikely to have an impact on the total volume of trading.

However, this order of magnitude change does seem, at least in the USA, to change the pattern of trading. A study by ITG suggests that the significantly smaller tick size (measured in basis points) increases the proportion of trading that takes place on traditional exchanges (NYSE, including ARCA and NASDAQ) and reduces the number taking place over the counter.[57] Measured in terms of visible liquidity, the cost of immediacy for small orders (below transaction sizes of around $3m) falls, but for large orders above this level it rises.[58]

[56] FESE (2011), 'Updated FESE tick size tables as of September 2011', September, available at http://www.fese.be/_lib/files/UPDATED_FESE_TICK_SIZE_TABLES_AS_OF_SEPT_2011.pdf

[57] See Table 1 in ITG (2011), ‘Trading Patterns, Liquidity, and the Citigroup Split’, August.

[58] Ibid., see Figure 1.

In relation to high-frequency activity, although it is not possible to identify the precise impact, the number of cancellations observed increases (by about a factor of 2) and in particular the number of fleeting cancellations (orders cancelled within 200 milliseconds) increases (by about a factor of 7). However, the ratio of trades per cancellation only declines marginally (from 0.087 to 0.075).[59]

Other common stocks that have experienced similar reverse stock splits in the USA have exhibited some of the same characteristics. Trading has generally moved towards exchanges and away from other trading venues (including OTC). The number of observable cancellations increases, particularly fleeting cancellations, but the ratio of trades per cancellation is much more stable.[60]

Although the reduction in tick size does seem to have an impact on some high-frequency metrics, which can be seen as an increase in high-frequency activity, when measured relative to the volume of transactions on the trading venue, the impact is limited. In addition, the changes in tick sizes (measured in basis points) that generate these impacts are much more significant than those that would result from regulatory intervention as envisaged in Europe. This evidence suggests that the impact of simply increasing the tick size marginally would be unlikely to have any significant impact on HFT activity.

5.4 Impact on trading sequences

5.4.1 Market-making strategies

Where the spread is one or (possibly) two ticks (which characterises highly liquid securities) increasing the tick size will tend to increase the rewards of a market-making transaction across the spread. This in turn would (slightly) reduce the need to combine market making with a directional element, but is unlikely to displace this directional element from high-frequency market-making strategies. Each tick may have slightly more resting orders in it therefore slightly increasing the time taken (on average) to go from the back of the queue to the front of the queue in the tick at the touch.

However, the general strategy of high-frequency market-making of lining up balanced trades at the front of the queue, or executing the hard side of the round trip first, or combining the spread with a (predicted) price movement, will still require significant repositioning of orders to successfully execute the market-making strategies. Increasing the tick size is, therefore, unlikely to affect the high levels of order cancellation that arise from market-making strategies.

5.4.2 Directional strategies

Increasing the tick size will tend to reduce the frequency with which price changes occur (ie, the price at the touch moves from one tick to the next). However, although movements become slightly less frequent, each move is worth slightly more.

[59] Ibid., see Table 1. It would appear that the increase in the share of trading on traditional exchanges is partially compensating for the higher number of cancelled orders on those exchanges

[60] Ibid., see Tables 7 and 8.

Like market making, changing the tick size will not change the need for orders to be positioned in each tick to take advantage of (predicted) price changes as and when they are (predicted to) occur.

5.4.3 Arbitrage strategies

Simple

An increase in tick size will tend to increase the value of fleeting mispricing as the price in each venue momentarily moves apart. Where the mispricing is one tick (for example, when the price in the leader venue has moved, but the price in the follower venue has not) a bigger tick will increase the value of that mispricing (for any given volume of security). However, the slightly larger tick may also slightly reduce the frequency with which these pricing anomalies occur.

Where resting orders are going to be used in the trading sequence the same issues relating to positioning and repositioning of orders will occur. Hence, the issues around cancellation rates are unlikely to change.

Linked

The same considerations apply. The frequency with which mispricing occurs may decrease slightly, but the value of mispricing increases (slightly). The increased tick size is unlikely to alter the basic trading strategy and the need to position and reposition resting orders.

Statistical

The same considerations apply, but now in a much more complex set of transactions.

6 Minimum resting times

6.1 The proposed rule

The MiFID consultation paper proposed that:

> Market operators would be required to ensure that orders would rest on an order book for a minimum period before being cancelled.[61]

In the Commission’s near final version of its proposals, this had been reworded as follows:

> Impose minimum latency period of orders in the order book—Under this option an obligation would be implemented according to which orders on electronic platforms would need to rest on an order book for a minimum period of time before they can be withdrawn.[62]

[61] European Commission (2010), ‘Review of MiFID’, December 8th, p. 16.

[62] European Commission (2011), ‘Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments’, p. 70.

In the final set of proposals published by the Commission this requirement had been dropped. However, since there is still some degree of fluidity in the final set of interventions, the analysis that follows below may still be relevant.

6.2 Rationale behind the rule and the Commission's impact assessment

According to the Commission, the proposed rule would stop high-frequency traders and algorithmic traders from testing the depth of order books by submitting and cancelling orders in very quick succession. This would put less stress on the IT systems of market operators, reducing the risk of system failures. Although there are strategies that use orders and cancellations to 'test' the depth of the order book, the conditions under which orders that do not execute can expose hidden depth are limited.[63] In dark markets such small testing orders can be used to search for liquidity, but orders not executing (and, therefore, being cancelled) will only indicate that liquidity is not present.

This proposed rule may also be intended to address the more general concern that high cancellation rates result in too much 'noise' and may make it more difficult for conventional traders to operate in the market. For example, orders placed by high-frequency traders may give other traders the impression that there is a considerable amount of liquidity, whereas, in practice, multiple orders may have been placed, only one of which is meant to execute, so the apparent liquidity may quickly disappear before it can be executed against. Many HFT sequences that involve multiple legs and resting orders may have the second (or more) leg cancelled if the first leg fails to execute (for example, because that trader fails to be 'first').

At the same time, the rule may address the concern about high-frequency traders (or, indeed, non-high-frequency traders) sending orders to an exchange without any real intention for them to be filled (which is likely to be a breach of the existing rules).[64] There may be a concern about some high-frequency traders being involved in strategies such as layering and quote stuffing. However, to the extent that these strategies involve sending orders to a trading venue that are some distance from the current touch, the probability that they would actually execute within a short resting time is actually small. Hence the impact on this type of messaging flow is likely to be rather limited.

Finally, it has been argued that the rule could be used to curtail HFT more generally in order to prevent any potential negative effects—it would affect HFT at its source by imposing a speed limit on trading. The argument is that while imposing minimum resting time would raise the average bid–ask spread (and therefore impose a cost to the end-investor), it might also lower its variability at times of stress.[65] According to this argument, liquidity would on average be more expensive, but also more resilient.

[63] One example is sending a small limit order to a venue inside the touch (ie, an offer to sell at a lower price than is currently visible, or an offer to buy at a higher price than is visible) to try to establish if there are any hidden resting orders. If the order executes then such a hidden order is present. If it does not execute there are no hidden resting orders at that price or better.

[64] See, for example, the recent action taken by the FSA in relation to a trader (FSA/PN/075/2011, August 31st 2011).

[65] See, for example, Haldane, A.G., (2011), 'The Race to Zero', Speech given by Andrew G. Haldane, Executive Director, Financial Stability and member of the interim Financial Policy Committee, International Economic Association Sixteenth World Congress, Beijing, China, July 8th.

The Commission mentions that the proposed rule has certain disadvantages and concludes that the costs are likely to outweigh its benefits. This rule is therefore not one of the Commission's preferred options. It prefers to impose an order to executed transactions ratio (discussed in section 7). The Commission identifies the following 'disadvantages':

> A disadvantage would be that a minimum latency period would limit market liquidity and efficiency and price discovery. The ability to constantly update orders helps maintain a tight bid-ask spread. In so far as some automated trading practices can be abusive this is an issue that will be addressed in the review of the Market Abuse Directive. This option would also amount to a prohibition of many forms of algorithmic and high frequency trading strategies that are considered to be beneficial to the market (e.g. market making and arbitrage strategies) where constantly updating orders is essential to enable the firm to provide the best prices and mitigate its risk. In addition, this measure could also indiscriminately affect other forms of trading where it is necessary to cancel or update orders. It therefore has the potential to distort the functioning of the market and create various unintended consequences. Finally, defining the minimum period would be highly controversial and sophisticated market participants may find innovative ways to exploit this resting period to their advantage.[66]

6.3 Reducing stress on operators' IT systems

If the single purpose of the proposed rule is to prevent large numbers of orders from putting stress on IT systems of market operators, then it should be borne in mind that, in practice, market operators have increased their capacity over time to cope with additional volume. They have also imposed restrictions on their users in terms of the ratio of messages to executed orders, for example, if it is not efficient to further increase capacity.

In other words, it is not clear whether there is a need for regulators to impose minimum resting times. Market operators would already have the relevant incentives themselves to deal with this. The Commission's impact assessment itself provides an overview of current practice in relation to minimum order to execution ratios.

In addition, one of the main mechanisms by which a minimum resting time would reduce the number of messages is if the restriction on the ability to cancel an order very quickly would be likely to result in that order executing when it (now) should not. In most of the HFT strategies the orders that are very quickly cancelled are cancelled because they are no longer needed, because of the change in circumstances of the market and/or trader. This change in circumstances can lead to either a position where the order is unwanted and still *likely* to execute, or where the order is unwanted and now *unlikely* to execute. In these latter circumstances, the imposition of a small resting time requirement is unlikely to have much impact on the unwanted execution rate, and as a result, only a limited impact on the trading sequences that can be pursued (see below).

6.4 Are HFT strategies harmful/abusive?

Strategies using HFT can be categorised in terms of their potential direct impact on other market participants, market confidence and end-investors, as follows.

[66] European Commission (2011), 'Commission staff working paper, Impact Assessement accompanying the document: Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p.128

- *Negative impact unlikely*—strategies that are unlikely to be directly harmful and may improve the price-formation process, such as market making and arbitrage strategies. All else equal, by improving the price-formation process, such strategies are likely to improve market efficiency.

- *Negative impact likely*—strategies that are likely to be directly harmful, such as quote stuffing, layering (at least when used to try to create a false impression of state of the order book) and, possibly, pinging orders.[67] These strategies are not necessarily new, but in some cases may have become more effective, inexpensive and attractive when they are implemented using HFT. Most, if not all, of these types of strategy are already banned or constrained under current regulatory frameworks. There is little understanding of the extent to which these trading practices are used by high-frequency traders. From a policy perspective, the question is whether HFT has meant that these abusive trading strategies have become more widespread, and whether it has become more difficult to detect them.

- *Impact or scale of impact unclear*—traders may be able to detect changes in the short-term supply and demand for securities from the analysis of the recent trading data (eg, concluding from recent market data that a large sell order is being put through the market, and then predicting that the price will therefore fall in the short term), and then trade ahead of it in anticipation of how the price will change (a strategy that is akin to front running). In most markets front running is illegal when the front-runner *improperly* obtains information about the incoming order—eg, when a broker obtains information about clients' orders. In the case of HFT, this type of front running can be based on information that is detected by analysing publicly available order data. In principle, this type of front running is not illegal, since the data is available to all market participants (at least in theory).

In the case of a predicted price increase, one strategy is immediately to submit a buy order that will execute by lifting an existing resting sell order (eg, market order). This step will increase the anticipated buying pressure, thereby reinforcing the predicted price movement. At the same time, anticipating the price movement, the high-frequency trader places one or more sell limit orders that are further up in the order book. By closely monitoring the subsequent order execution and other changes in the order book as the price rises, the high-frequency trader can attempt to allow the optimal resting order to be filled (for example, by cancelling, before execution, resting orders in the order book that are not at the level, the predicted price movement will run out of steam).

The high-frequency trader will benefit from this strategy if the price difference between the first-order buy order and the subsequent filled sell order is larger than the costs of trading (eg, exchange fees, clearing fees, costs of other inputs). All else equal, and on the assumption that the trading decisions of other market participants are not altered over the time period when this HFT strategy executes, the final holding position of the other market participants is the same (the high-frequency trader has returned to a flat position) and the net purchasers have paid slightly more for their securities (by the gross amount the high-frequency trader has cleared from its trades). The number of transactions has also increased.

The high-frequency trader could, however, initiate its trading strategy not through a buy order that executes against an existing resting sell order, but by placing a resting buy order one tick ahead of the existing resting buy orders. This order would then execute against the next

[67] For an explanation of abusive strategies, see AFM (2010), op. cit.

incoming sell market order (or equivalent). In this case, the subsequent part of the strategy can be the same. All else equal, the impact on other market participants is more complex. The incoming sell market order now executes at a better price for the seller than it would otherwise have done. An existing buy resting order at the old touch does not execute. There are no additional buying orders against existing resting sell orders, and an additional resting sell order (that executes) is added to the order book (probably along with other sell orders that are cancelled before they can execute). The overall impact is that an existing resting buy order at the old touch does not execute and an existing resting sell order some way down the order book does not execute. If these two orders that do not now execute have resting orders supplied by market-makers then, as a class, these market-makers are worse off than they would otherwise have been. The profit on the trading sequence made by the high-frequency trader, and any price improvements experienced by those sending executable trades to the market, are ‘paid’ for by the reduction in profits experienced by the other market-makers. Within the period of the trading sequence, the effective spread experienced by executing market orders (or their equivalent) has reduced. (If, however, the resting orders that do not execute have been placed by end-investors, as a class it is they that suffer this reduction in profit.)

In practice, the trading sequences are likely to be more complex, and may well alter the behaviour of other market participants. However, what the stylised examples above illustrate is that the impact on other market participants as a result of the high-frequency traders exploiting short-term predicted price movements is likely to depend critically on the precise trading strategy they adopt. A single or uniform impact is unlikely to arise from the activities of high-frequency traders when they trade using their predictions with respect to short-term price movements.

It has been argued that ‘executing large orders without moving the market is an utopia’, and that ‘the fact an investor is buying a large amount of shares obviously impacts demand and should, in an efficient market, trigger a rise in the stock price.’[68] This problem of the market impact of large orders is not new. Given the ambiguous impact on those wishing to achieve large net transfers of securities, direct empirical evidence on market impact is likely to be needed to fully understand the impact of HFT when undertaking this type of trading strategy.

In addition, other developments in the market place will be having an impact on the market. Traders with large orders can nowadays attempt to avoid front-running strategies of those trading against them by using ‘safe havens’, such as delayed trade reporting or using dark pools. However, it has been argued that, even in a dark pool, depending on the rules, front running may occur.[69]

To assess whether these types of trading strategy significantly affect investors would first require an empirical analysis to understand to what extent such strategies are being pursued. Some empirical studies have started to explore these issues. For example, a recent academic paper describes an analysis of high-frequency traders’ trading data that is used to assess to what extent high-frequency traders are involved in front-running strategies.[70] The analysis suggests that front running by high-frequency traders before large orders is not systematically occurring.

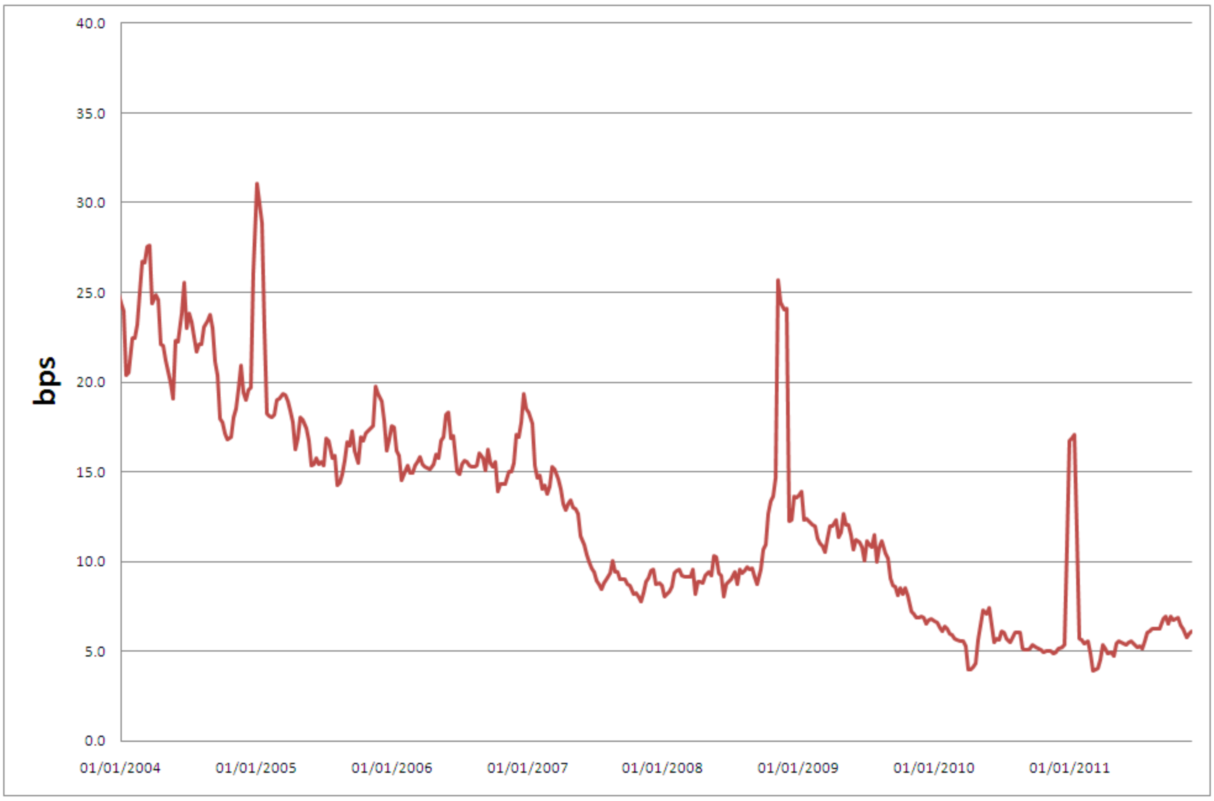
[68] See, for example, Optiver (2010), op. cit.

[69] See, for example, *Wall Street Journal* (2008), ‘Trading in a Dark Pool? Watch for Sharks’, August 18th.

[70] Brogaard, J.A. (2010), ‘High-frequency trading and its impact on market quality’, Northwestern University, July 16th.

Another way to approach this is to assess the extent to which bid ask spreads and market impact have changed over time. Using publicly available data that has been collected for other purposes (and must therefore be interpreted with caution), we can observe that for the FTSE 100 shares the bid ask spread has been generally falling for the period 2004 to at least 2012, after which it may be flattening out (see Figure 6.1) while there does not seem to be any significant worsening, or improvement, in market impact over the past few years, during which time the level of HFT has increased (see Figure 6.2).

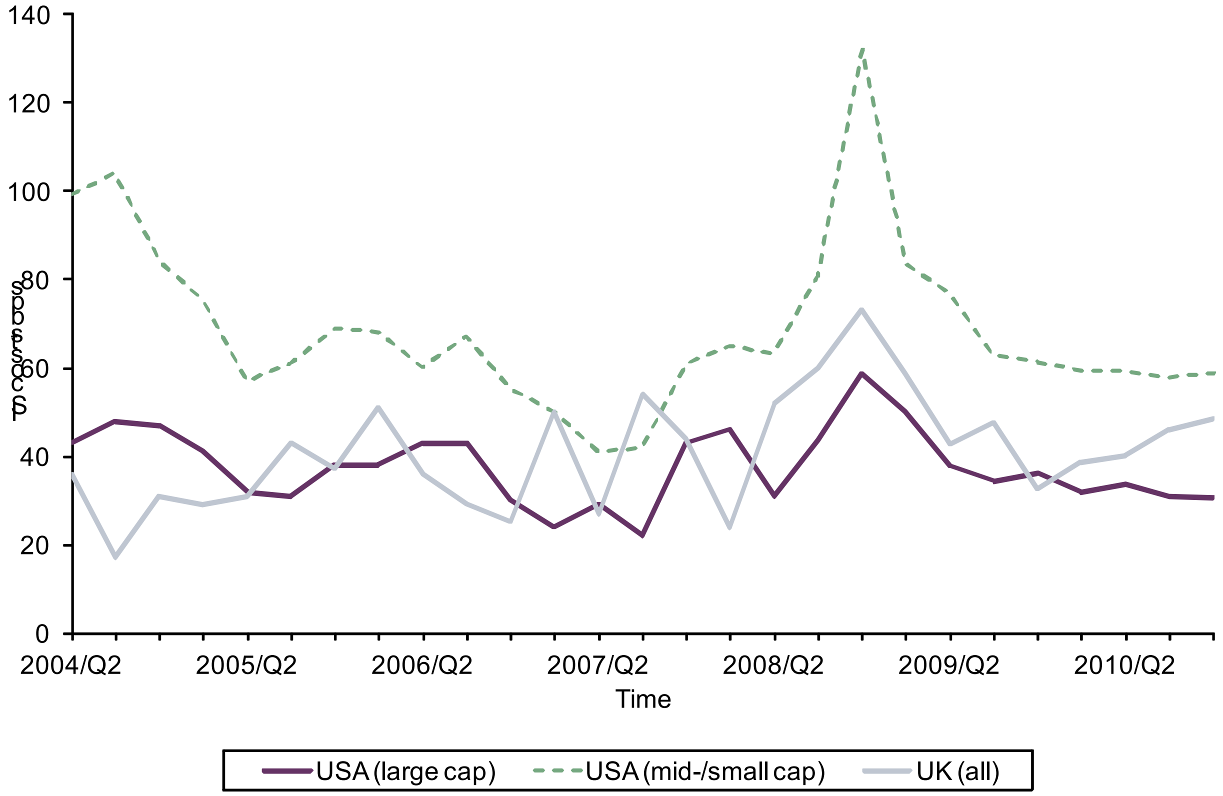
Figure 6.1 Changes in the bid ask spread for FSTE 100 shares



Note: IS defined as: 'the difference, or slippage, between the arrival price and the average execution price for a trade'.
Source: Investment Technology Group (2011 and 2008, various dates), 'ITG's Global Cost Review', http://itg.com/news_events/papers/ITGGlobalCostReview_2010Q3_Final.pdf.

Figure 6.2 shows the market impact measured from an institutional investor perspective. As can be seen, the market impact being recorded now is similar to that recorded before the financial crisis, and this is similar for both US large cap stocks and US small cap stocks, while HFT is much less prevalent in the trading of small cap stocks. As indicated, the IS dataset is not ideal for the analysis that would ideally be undertaken here, and the results should therefore be treated with caution. However, the absence of any large-scale trend on market impact (at least measured in this particular form), the general reduction in spread (at least until recently)combined with the theoretical conclusion that the impact of HFT is likely to be ambiguous, and the indication from the interviews that there are instances of 'negative' impact, but that the overall impact is much less clear, all tend to reinforce the premise that, if there are negative consequences on market participants, these will vary both by the precise trading strategies undertaken by the high-frequency traders and the trading strategies used by the other market participants.

Figure 6.2 Implementation shortfalls costs in the UK and USA (basis points)



Although it cannot be ruled out that certain traders implement certain strategies (for example, layering and quote stuffing), there is no evidence that this is taking place at present on a larger scale than was occurring before the rise in HFT—an empirical analysis would be required to determine this.[71]

If the Commission were concerned about abusive practices such as layering and quote stuffing, then it would be more appropriate to first explore the possibility of closer scrutiny and market surveillance through transaction reporting by investment firms, traders and trading platforms. The disadvantage of imposing minimum resting times is that all HFT (and to some extent also non-HFT) would be affected rather than just those strategies or practices that are harmful.

6.5 Why are cancellation ratios so high?

The relatively high cancellation rates can be, at least partly, explained by other factors, such as trading strategies that do not necessarily involve abusive practices but have other rationales.

- Search for liquidity through pinging—the SEC notes that ‘there is an important distinction between using tools such as pinging orders as part of a normal search for liquidity with which to trade and using such tools to detect and trade in front of large trading interest as part of an “order anticipation” trading strategy’ (which is referred to as front running).

[71] Some dark pools impose minimum resting times to some extent, to prevent traders from submitting and cancelling orders to test the liquidity in the dark pool.

- HFT strategies are likely to be probabilistic in terms of both whether the trading strategy will be successful and whether it will actually be executed in a particular instance. Strategies that require a sequence of executions (for example, arbitrage strategies, directional strategies) may fail at the first point, in which case all subsequent orders need to be cancelled before they can execute. When one party submits identical orders to multiple venues, execution in one venue may require cancellation of the identical orders in the other venues.

- High-frequency traders are likely to be constrained in capital and risk. They may send many different sequences of orders that they would like to be executed, but, in practice, it is possible to execute only a subset of these orders at any one time and remain within the trader's risk and capital appetite. For example, if a high-frequency trader de facto acting as a market-maker approaches its position limit, many or all the remaining outstanding orders may need to be cancelled until such time as the hit or lifted orders are reversed, thereby returning the HFT market-maker to a flat position, and thus allowing it to place more orders, some of which are again hit or lifted. This would tend to generate high cancellation rates when the probability of resting orders being filled in any time period is low. In other words, a large number of bids and offers are needed to ensure that some orders are executed; however, when they are executed other remaining orders may need to be cancelled, until the round trips of the executed orders are completed.

- Directional strategies may involve submitting multiple orders, each of which would return the high-frequency trader to a flat position, at different prices within the order book, with the optimal order to be filled chosen within the holding period. Orders at prices that appear to be improbable need to be cancelled before they can execute, and orders that are predicted not to be filled will need to be cancelled once the flattening order has been executed.

The cancellation rate or hit rate (defined as the percentage of orders sent that are actually executed) can vary considerably by trader. Anecdotal evidence suggests that hit rates in relation to HFT market-making strategies are much lower (potentially less than 1%) than, for example, the hit rates in the case of other trading strategies implemented using HFT. This is because HFT market-makers continuously track the movement of market prices by updating their limit orders.

6.6 General impact

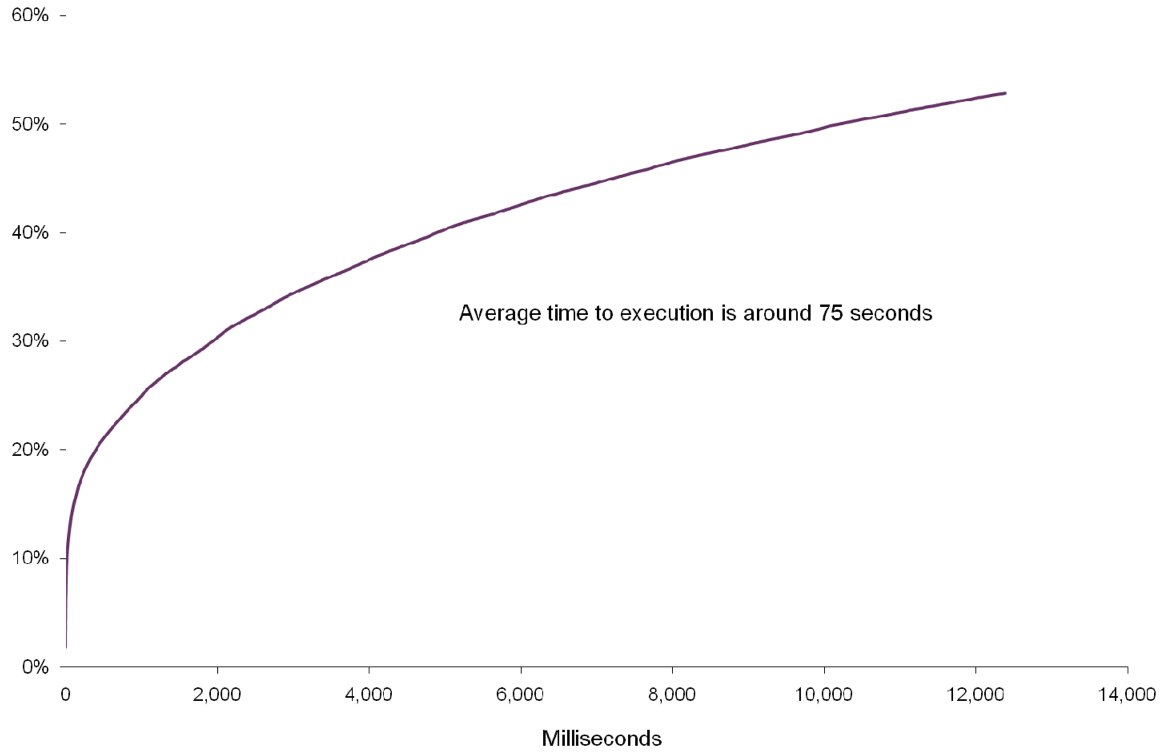
A major characteristic of HFT is the number of orders that are sent to the trading venue and that are cancelled very quickly. A minimum resting time requirement would not allow these orders to be made in this form and, therefore, could have a significant impact on the conduct of HFT.

However, at a general level, it is only *if* those orders that are cancelled very quickly (ie, within the time limit) would *actually* execute if left for the period to the time limit, will there be any significant impact on HFT. If all that happens is that these orders remain on the order book for the minimum period and are then cancelled, that the impact on the economics of HFT is minimal.

As a result, the impact of a rule of this sort arises from orders that are sent to the trading venue, fail to execute immediately, are now no longer wanted, but then execute within the minimum time period. This is quite a restrictive set of conditions within which the impact will arise, and does not cover a significant part of those orders that are currently sent to a trading venue and cancelled if they do not execute immediately.

Data for 2009 for ten major securities on the NASDAQ exchange suggests that for resting orders that *do* execute, the average time that the order has been resting for is 75 seconds, and the median time is ten seconds. Figure 6.3 below shows the cumulative frequency of resting times for these executing orders (up to 12 seconds). However, notwithstanding the ten-second median and 75-second average, a significant proportion of orders (~10%) that do not execute on arrival do execute very quickly after they have arrived (in under 20 milliseconds).

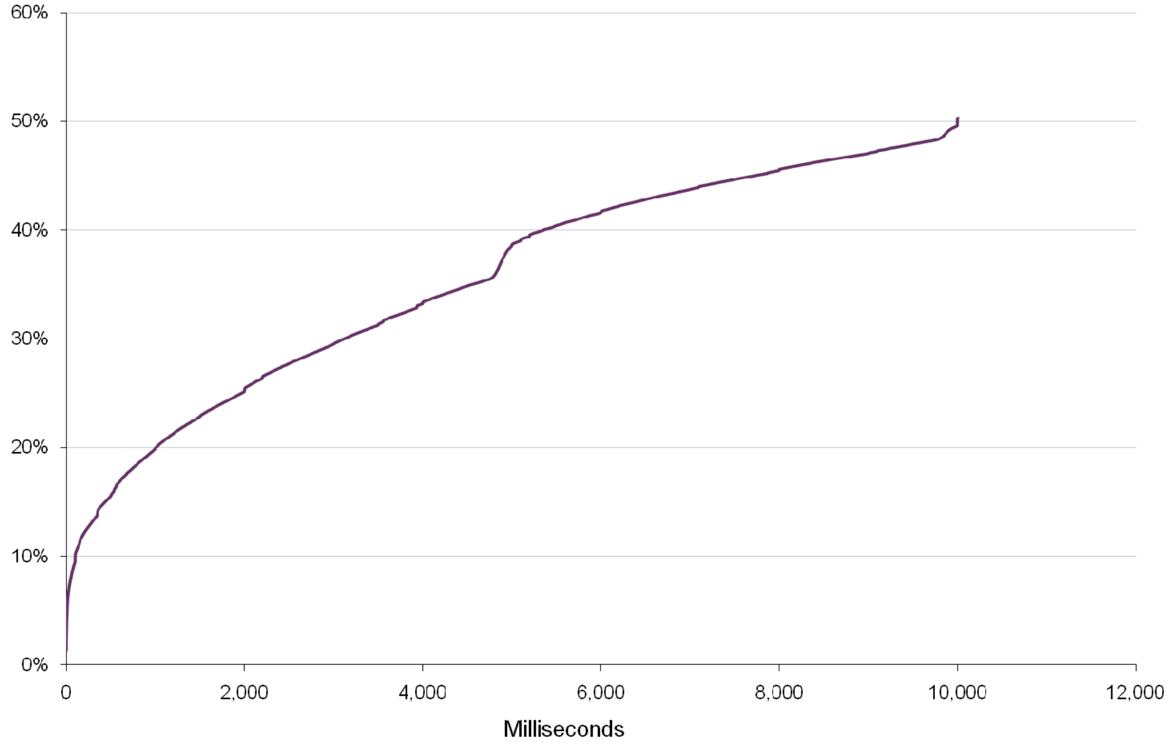
Figure 6.3 Cumulative frequency distribution of resting time to execution (ten securities on NASDAQ 2009)



Source: Nikolaus Hautsch and Ruihong Huang (2011), ‘Limit order flow, market impact and optimal order sizes: evidence from NASDAQ TotalView-ITCH data’, and Oxera analysis.

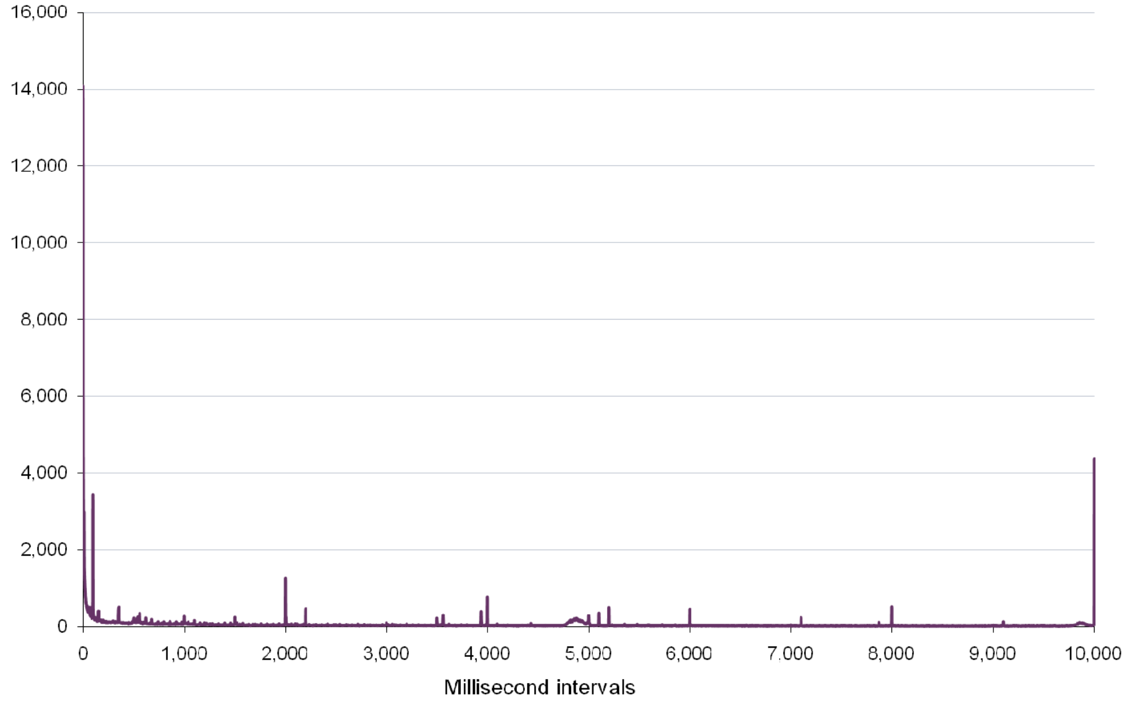
In addition there are significant levels of orders that are cancelled quickly. For the same set of securities, the time to cancellation is set out in Figures 6.4 and 6.5.

Figure 6.4 Cumulative distribution of resting time to cancellation, up to ten seconds (ten securities on NASDAQ 2009)



Source: Nikolaus Hautsch and Ruihong Huang (2011), ‘Limit order flow, market impact and optimal order sizes: evidence from NASDAQ TotalView-ITCH data’, and Oxera analysis.

Figure 6.5 Frequency distribution of time to cancel—nominal count, millisecond intervals, up to ten seconds (ten securities on NASDAQ 2009)



Source: Nikolaus Hautsch and Ruihong Huang (2011), 'Limit order flow, market impact and optimal order sizes: evidence from NASDAQ TotalView-ITCH data', and Oxera analysis..

These figures show that a significant (~6–7%) of cancellations take place within 20 milliseconds and 12% within 200 milliseconds (15% within 0.5 seconds). However, they also show that there is a considerable number of cancellations that would appear to be programmed to take place at rather arbitrary times from a market dynamics perspective—ie, at two seconds, four seconds, five seconds etc, as well as at ten milliseconds, 100 milliseconds, etc. This suggests that in many cases the risks of executing at the wrong time are not carefully considered and that it would, therefore, be a mistake to conclude that all cancellations are because if those orders remained on the market they would be at a serious risk of executing. Rather than being cancelled because they are likely to execute at the wrong time, these orders could be being cancelled simply because they are no longer needed. Indeed, one of criticisms that has been made in relation to HFT is that orders are placed and cancelled in very quick succession at positions in the order book where such orders would have a very low probability of execution. Although these orders will be captured in an analysis of the cancellation times, if they are a long way from the touch the probability of them executing within a minimum resting time would be (very) small.

Although there are a large number of order types that can be sent to trading venues, where there is a price–time priority system (ie, orders that rest on the exchange are executed in order of price then within price by time of arrival) there are a limited number of places within the order book that an order can be placed. In essence, there are three positions, as follows.

1. The arriving order can be executed against an already resting order at the venue. Such an order executes on arrival, and cannot be cancelled, no matter how fast the trader.

2. The arriving order is at a price at which there are other resting orders. This order is placed at the back of the queue in that tick. At a minimum, such an order will only execute once all the orders ahead of it in that tick are either cancelled or executed. In addition, any orders arriving at the exchange at a better price will execute ahead of the relevant order. At any time before these conditions arise, this order can be cancelled without executing.

3. The arriving order represents a better price (from the perspective of the counterparty) than any existing order resting on the exchange, but is not so good that it can execute against any existing resting order. Such an order will be placed at the head of an empty tick which will represent (at that moment) the best price available. This order will execute when the next suitable marketable order arrives (ie, if the relevant order is a sell order, when the next marketable order to buy that security arrives), unless in the mean time a new order at an even better price has arrived, and is placed in another empty tick.

A minimum resting time is not relevant to orders of the first sort. The minimum resting time will bite on orders of the second sort if the resting time is longer than the period taken for all the orders in front of the relevant order to be cancelled or executed. This time is clearly dependent on the number/volume of orders ahead of the relevant order and the speed at which these are being cancelled or executed. In 2011, the average frequency of execution of a security on the LSE in the FTSE 100 was around one bargain every eight seconds (ie, one buy and one sell), although this will clearly vary by time of day and security. From the perspective of an order resting on one or other side of the spread, this is an execution frequency of one every 16 seconds on each side. The most heavily traded securities will clearly have much more frequent transactions, but even Vodafone and BT have between 200,000 and 300,000 bargains per month (Vodafone) and 80,000 and 120,000 (BT) (in 2011), which only increases the frequency around one every two to three seconds (Vodafone) or one every five to seven seconds (BT) on average.[72] Even if trading is concentrated within the day, the number of transactions on each side of the spread per second in this high volume period for most securities on the LSE will be a relatively small number, if not less than one. If there are multiple orders already resting at the touch when an order is placed at the back of the queue (ie, a type 2 order in the above list) its average time to execution will tend to be seconds, not milliseconds. This raises the possibility that, at least in normal circumstances, relatively short minimum resting times (for example, 200 milliseconds) would not have a significant impact on type 2 orders executing at the wrong time.

Even in less normal times, since part of the HFT approach is to closely monitor the state of the order book and to predict (short-term) price movements, it may be possible to devise algorithms that would be able to predict the risk of an order that does not execute on arrival or very quickly (ie, executing when it was supposed to execute for the trading strategy to be successfully implemented) and the risk of executing at the wrong time within the minimum resting period window. Such a prediction would tend to eliminate some high-risk orders, but would be unlikely to result in a significant proportion of these types of orders being eliminated.

The analysis above is based on an assumption that trading in a particular security is spread reasonably throughout the trading period when analysed over very short time periods. That is,

[72] In 2011 the LSE transacted just over 100m bargains (one buy and one sell) in FTSE 100 securities, or around 1m bargains per security. With trading taking place on 251 days, this is around 4,000 per day per security (on average). The trading day is 8.00am to 4.30pm, which is 30,600 seconds. On average, therefore, one bargain per security takes place every 7.5 seconds. For high-volume stock—for example, Vodafone—the number of trades per month on the LSE ranges around 200,000–300,000, or 10,000–15,000 per day. This represents one bargain every two to three seconds. See http://www.londonstockexchange.com/exchange/prices-and-markets/stocks/exchange-insight/trade-data.html?fourWayKey=GB00B16GWD56GBGBXSET0, and annual trading statistics from the LSE.

that trading is not concentrated into a relatively small number of very small time windows. However, if this were to be the case, then at time periods when trading is actually taking place the trading frequency could be very much higher than the analysis set out above would suggest and that, in these time periods the elapsed time between placing an order at the back of the queue at the touch would execute very quickly because the orders ahead of it would themselves be executing quickly because the frequency of execution would be very high.

Empirical analysis of trading frequency and volumes in the tick at the touch at a very granular level should be able to establish if this is the typical pattern (or, more realistically, the proportion of trading that occurs under these conditions). If this pattern does represent a high proportion of trading of individual securities then the impact of minimum resting periods would be more significant, notwithstanding the average frequencies of trading that suggest that most orders (excluding those sent within the touch—type 3) tend to rest for quite long periods.

However, even under these circumstances, the response from high-frequency traders is likely to be to stop placing resting orders in the relevant ticks when very short resting periods are predicted, rather than placing orders that have a high probability of executing at the wrong time.

The third type of order is much more likely to execute very quickly. Analysis of one security on NASDAQ in 2009 suggests that the average time to execution of an order that is sent to rest inside the existing spread is only 400 milliseconds.[73] These types of orders are expected to execute against two types of incoming order that arrive at a later time. The first is a marketable order that arrives independently of the placing of the relevant type 3 order, which has now become the best quote and is resting in the book. In other words, the arrival of the marketable order was not triggered by the new type 3 order becoming visible to other market participants.

The second type of order is itself triggered by the visibility of the relevant type 3 order, and results in at least one matching order being sent to the exchange within the minimum resting time window.

In the first case the probability of a marketable order arriving and executing against the recently posted order within the minimum resting time window will be determined by the general (marketable) order flow to the venue. This will essentially match the frequency of transactions. For instance, if these marketable orders arrive once every five seconds on each side, 4% would arrive within 200 milliseconds of the type 3 order arriving. The risk of execution from this source would, therefore, be real, even if relatively small.

However, the second type of order (the triggered order) may represent more of a threat and will occur if the minimum time window is long enough for a trader who wishes to execute a trade within the touch to recognise the availability of the resting order and to send the matching order to the trading venue. The minimum response time will be twice the transmission time from the trader to the venue, the time taken for the trader's computers to recognise the new order and to dispatch the matching order. (There may also be some processing time within the trading venue's computers depending on where the time window is measured.) For co-located traders with fast processing, the minimum resting time window will be likely to be longer than the processing and communications time. As a result, an assumption can be made that type 3

[73] Nikolaus Hautsch and Ruihong Huang (2011), 'Limit order flow, market impact and optimal order sizes: evidence from NASDAQ TotalView-ITCH data'.

orders will generally be vulnerable to execution within the window from trades generated as a result of the exposure to the market of the relevant order.

As a result, if orders sent inside the touch that fail to execute on arrival would then execute at the wrong time, a minimum resting time would have a very significant impact on these orders. Since these resting orders will represent the best price on offer (at least immediately after they arrive) the probability that they would be attractive to some other trader is likely to be high, and it would be expected that most of these orders would tend to execute in the presence of a minimum resting time. For this type of order, therefore, an assumption can be made that minimum resting periods will have a significant impact on the execution rate and hence have the possibility of affecting the behaviour of those who are sending such orders.

6.6.1 Impact on the costs of immediacy

As explained above, with the exception of type 3 orders, there will be a time interval between an order arriving at the back of the queue and it being available for execution at the front of the queue, and that a minimum resting period will bite only to the extent that this interval is less than the resting period. Much more detailed empirical analysis of the distribution of this time interval would be necessary to gain an understanding of the relationship between the value chosen as the minimum resting period and the impact on trading sequences, and hence the impact on HFT. In addition, an analysis of the predictability of this time interval in relation to the placing of an individual resting order would also be necessary to capture the likely response of HF traders.

If the individual time intervals (ie the time taken between arrival at the back of the queue at the touch and reaching the head of the queue) are highly predictable HF traders can be expected to continue to place orders at the back of the queue where the prediction indicates a high probability of the time interval being larger than the minimum resting period, and not to place orders at the back of the queue at the touch when the prediction is that the time interval will be less than the minimum resting time. In addition, where the time interval is not very predictable orders again will tend not be placed at the back of the queue at the touch.

Placing orders at the back of the queue in ticks away from the touch will, by definition, place more orders ahead of the relevant order and, therefore, increase the probability of the time to possible execution interval being longer than the minimum resting period.

As a result, HF traders can be expected to reduce the placing of orders at the back of queue in the tick at the touch, and to increase the placing of orders in ticks away from the touch (or what would have been the touch in the absence of the minimum resting period rule). In addition, in those periods of the trading day where there is a relatively rapid execution of orders this behaviour response by HF traders can be expected to be more pronounced.

Generally, therefore, as the minimum resting period rule bites (ie as it is made longer) HF traders would tend to place fewer (if any) orders at the touch (or what would have been the touch) and to potentially place more pre-positioning orders away from the touch. Observed spreads would, therefore, tend to widen.

Although the relationship between widened spreads and the price of immediacy is not completely straightforward (especially for large investor orders) they will tend to translate into a higher price for immediacy, particularly for smaller investor orders. Because of the volume of trading, a small increase in the price of immediacy can have a large total impact. In order of magnitude terms, if the (approximate) market capitalisation of the FTSE 100 (£1.5tr) is taken as the proxy for the market capitalisation of all European equity securities being currently traded in a significant way by HF traders, and the investor turnover of those securities purchased or sold with immediacy is taken as 0.5, a one tick increase (~5bp) in the effective spread would result in an increase in the cost of immediacy of around £375m.[74] It is difficult to exactly predict the impact of an increase in resting times on the bid ask spread but this analysis suggests that if the increase in resting times were large, the impact in terms of the increase in the costs of immediacy could be billions rather than millions.

6.7 Impact on trading sequences

6.7.1 Market-making strategies

Market-making strategies that are using pre-positioned (resting) orders that are allowed to execute when the prediction is that a trading sequence initiated at that time will be profitable will tend to be type 2 orders (as defined above). Hence the impact of a minimum resting time requirement may be muted as it will only bite when the minimum time is longer than the time taken to exhaust the resting orders already in the relevant tick.

Although it will clearly be impossible to predict the precise time interval between an order arriving at the back of the queue and that order reaching the front of the queue and executing, with full visibility of the order book it should enable high-frequency traders to be able to predict when the risk of execution before the minimum time period has expired is high, and when it is low. As a result, high-frequency market-makers should be able to adjust their behaviour accordingly and only place orders into a tick when the risk of execution at the wrong time as a result of the minimum resting time is low.

At present there appears to be no existing analysis of the behaviour of significant order books to enable a robust empirical estimate to be made as to what proportion of current type 2 orders would be at risk. However, data that suggests that the median resting time of all executed orders (ie, type 2 and type 3, as defined above) is in the region of ten seconds (on NASDAQ) is not inconsistent with the ‘at risk’ proportion of type 2 orders being quite small. Further detailed analysis of a major European order book could be undertaken to try to quantify more accurately this proportion.

Market-making strategies may also use type 3 orders where the trader wants to execute against the next marketable order but is not at (or near enough to) the head of the queue at the (existing) touch. The objective here will be to execute very quickly, but there may be execution sequences where the imposition of a minimum resting time will have an impact on the approach taken.

In particular, when an order has been sent for the same security to more than one venue, and the trader only wants one of these multiple orders to execute to initiate or to complete the market-making sequence. The imposition of a minimum resting time on each of these orders would significantly increase the risk of more than one executing. As a result, multiple orders to different venues within the spread would be likely to reduce. (Note, however, that this relates to type 3 orders only. Multiple type 2 orders would still be useable as long as the predicted time

[74] The equity market value of the companies with a market value of more than £2,000m amounts to £1,677,784,534,351 (see Table 8 in http://www.londonstockexchange.com/statistics/markets/main-market/main-market.htm).

interval from the back of the queue to the front of the queue in the tick is longer than the minimum resting period.)

In addition, it is possible that a high-frequency trader would want to use a type 3 order to complete the second leg of a market-making round trip (for example, because the resting type 2 orders that could also complete the sequence are too far from the front of the queue). However, under these circumstances it is likely that the trader actually wants the order to execute, so would leave the order in the market even in the absence of a minimum resting time requirement. A possible exception could be where the probability of the type 2 order executing (which would represent a better outcome for the high-frequency trader) may change within the minimum resting period and, therefore, the high-frequency trader would wish to cancel the type 3 order within this period. Again, it is unclear from the empirical analysis how frequent this occurrence is, but, in any case, even when the minimum resting rule bites the outcome is still one that the high-frequency trader was prepared to accept very shortly before.

Some impact on trading sequences could be expected, but for the reasons set out above, in relation to market-making sequences a minimum resting requirement may have a rather limited impact.

6.7.2 Directional strategies

To the extent that type 2 orders are used in these strategies the same conditions apply as described under market-making strategies above. Because directional strategies are predicated on the movement of prices through the order book, it will be possible for the second leg of the sequence to have been resting for some time. Indeed, the execution of these sequences is likely to be using orders that have been placed into these ticks when they have been one or two ticks away from the touch, and will, in nearly all normal market conditions, have been resting for some time before there is a risk that they will execute.

The first leg of a directional sequence will generally be using resting orders at the existing touch (or at a slightly worse price—from the perspective of the high-frequency trader—using a type 3 order). For the reasons set out above in relation to the market-making strategy these orders will be placed at the back of the queue and the risk of executing does not occur until they reach the front.

However, there clearly will be occasions when an order placed at the back of the queue in the tick at the touch will execute within the resting period. Under these circumstances the minimum resting time will bite, if the directional strategy does not want that order to execute. The circumstances when this occurs are likely to be when the price is moving in the opposite direction to that which constitutes the first leg of the directional strategy. However, to the extent that the high-frequency trader has predicted this price movement before placing the order at the back of the queue, the trader will be able to reduce the occurrence of these orders executing at the wrong time.

The risk facing the directional strategy from a minimum resting time can, therefore, be seen as the risk that the predicted direction of short-term price movement changes within the resting period combined with the time taken to move from the front to the back of the queue.

It should also be noted that the high-frequency trader already faces a similar type of risk with no minimum resting period. This arises because, if, after an order has been placed at the back of the queue a single marketable order is sent to the trading venue that is of sufficient size that it executes against all the existing resting orders in that tick, even the highest speed trader cannot get ahead of that order to cancel it.

To the extent that the interaction of the minimum resting time, the risk of predicted price direction change and the time taken to run down the tick at the touch interact, this will increase the risks faced by high-frequency traders using this strategy. However, the type of predictions they would need to make to reduce this risk would seem to be similar to the predictions they are already making. High-frequency algorithms are, therefore, likely to be able to take into account reasonably efficiently any minimum resting period (unless, of course, the minimum is very long).

The first leg of a directional strategy may also be a type 3 order, placed inside the touch. This order is very likely to execute in very short time scales. Hence a minimum resting time would have a significant potential impact. However, a type 3 order is likely to be used when the predicted price movement means that the trader's resting order at the touch will not execute because it would not reach the front of the queue while that tick remains at the touch. The high-frequency trader wants the type 3 order to execute if their price movement prediction is correct. If the prediction is correct, and the type 3 order does not execute very quickly (as wanted by the trader) this will be because either there are no matching marketable orders coming into the trading venue, or another trader has placed a resting order inside the relevant type 3 order. So to the extent that the prediction is correct, but the type 3 order does not execute, the risk that it will now execute at the 'wrong' time within the minimum resting period would appear to be very small.

However, if the prediction turns out to be wrong then the inability to cancel the order within the resting period does increase the probability of the order executing at the wrong time. The order will tend to represent the best price available and will be attractive to other counterparties because the price will tend to be moving in a direction that makes the price even better for those counterparties.

With minimum resting times, therefore, the high-frequency trader contemplating a type 3 order to initiate a directional sequence faces the additional risk that the prediction upon which the sequence is predicated is wrong, (strictly speaking it could be right but reverses within the minimum resting period). Under these circumstances the order fails to execute in the time frame within which the trader wishes the order to be exposed, but does execute between that time and the minimum resting period.

It is unclear from any empirical analysis how frequently these conditions would be met, and this type of risk does not seem to be fundamentally different from those which the high-frequency trader would be including in the current algorithms. This would suggest that this additional risk could be relatively easily factored into future algorithms.

Directional strategies may also be carried out across different venues. The same considerations as set out under market making will apply. Minimum resting periods will increase the risks of multiple orders executing when only one execution is required. For example, if two type 3 orders are sent to different venues the second order would remain exposed for the remainder of the minimum resting time after the first order had executed (and the second order is no longer wanted). But where resting orders have been placed in the ticks in anticipation that they might be needed in the future, even multiple venue orders will tend to have already been resting sufficiently when the need to cancel them becomes acute.

All of this suggests that although there are conditions under which directional strategies would be affected by minimum resting times, these would tend to arise under rather restrictive conditions, and conditions that could be included in the trading strategies and algorithms.

6.7.3 Arbitrage strategies

The same general considerations apply. There are some circumstances under which a minimum resting time will have an impact. These can be summarised as when:

- multiple type 3 orders are sent to different trading venues, but only one is actually wanted to be executed;
- predictions about changes in market conditions are made, triggering new orders, but these predictions turn out to be wrong and this becomes apparent within the minimum resting time, and the order did not execute prior to the error being spotted;
- the available liquidity in the tick at the touch is small and the turnover of the security is high, such that the time taken for an order to go from the back of the queue to the front of the queue is generally smaller than the minimum resting period.

6.7.4 Pinging

The activity of pinging is vulnerable to minimum resting periods, when the objective of the ping order is to establish that there are no hidden orders/liquidity inside the touch. An order that fails to execute on arrival (ie, establishes that there is, or appears to be, no hidden liquidity inside the touch) will then represent the best price available. Unless the fact that no hidden liquidity is available feeds into a strategy that needs that, now resting, order to execute, the execution of that order will represent an execution at the wrong time.

For example, if a security currently has the touch (ie, visible bid and offers) at 97 (offers to buy) and 100 (offers to sell), a pinging order to test for hidden liquidity could be a sell at 98 or a buy at 99. These will not execute if there are no hidden resting orders. A minimum resting period would make these orders visible to other traders who would have enough time to react and send marketable orders to the venue that would now execute at prices better than were previously available. If the absence of hidden orders is taken to be a neutral outcome in terms of predicting short-run price changes, the execution of these pinging orders now represents a worse price for the high-frequency trader compared with other resting orders that the trader has in the ticks at the touch (in this example, an offer to buy at 99 instead of an offer to buy at 97). Any market-making, directional or arbitrage trading sequence started from this ping order execution will, therefore, be less profitable than one executed from the earlier visible orders at (what was) the touch.

If the ping order does execute on arrival (ie, it does discover hidden liquidity) the implications will be that if a price movement is expected it will be in the opposite direction to that which would ideally be exploited by the ping order. In the example above, if the ping order of the offer to sell at 98 executes (ie, there is a hidden resting order of an offer to buy at 98) the implied price direction is upwards. This in turn suggests that the existing resting orders to offer to sell at 97 will not execute in the immediate short term. A high-frequency trader wishing to exploit the predicted movement would then need to send a visible offer to buy at 98 to step ahead of the resting orders in tick 97 and to get ahead of the hidden order already in place at 98. This is the opposite of the original ping order which executed (which was an order to sell at 98).

Ping orders being used to provide information for use in market-making, directional or arbitrage trading sequences will tend, therefore, to be such that if they do not execute on arrival (ie, identify the hidden liquidity) the high-frequency trader will not want them to execute. However, even very short minimum resting periods will be likely to increase the probability of this happening, and since these executions will generally be at the wrong time, it would be expected that the use of pinging orders would decrease markedly.

The reduction in the information available to HFT will in turn have some impact on the ability to make predictions on the future direction of prices, so that some reduction in trading sequence activity could be expected. However, the extent to which the information generated by pinging creates *relative* advantages between high-frequency traders, rather than an absolute ability to make short-term price predictions, the reduction in non-pinging trading sequences will be muted.

Pinging is also used by traders wishing to minimise the market impact of a large change in a net position. By identifying where there is, or is not, hidden liquidity, the trader can attempt to optimise the order placements at different venues so as to minimise market impact. However, in these cases, it is likely that the execution of a ping order as a result of a minimum resting requirement would represent a transaction that is actually wanted and would be executed at a better price than the visible orders at that time. A minimum resting period would be likely to have much less impact on the use of pinging orders for this type of activity.

7 Minimum execution ratios

7.1 The proposed rule

The MiFID consultation paper proposed an alternative to the minimum resting time rule:

> .... they would be required to ensure that the ratio of orders to transactions executed by any given participant would not exceed a specified level. In either case, further specification would be needed on the specific period or level.[75]

In the Commission's final version of its proposals, this has been worded as follows:

> Member States shall require a regulated market to have in place effective systems, procedures and arrangements to ensure that algorithmic trading systems cannot create or contribute to disorderly trading conditions on the market including systems to limit the ratio of unexecuted orders to transactions that may be entered into the system by a member or participant, to be able to slow down the flow of orders if there is a risk of its system capacity being reached and to limit the minimum tick size that may be executed on the market.[76]

The Commission has now come to the view that a minimum order to execution ratio would be preferred over imposing a minimum resting time.

[75] European Commission (2010), 'Review of MiFID', December 8th, p. 16.

[76] European Commission (2011), 'Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p. 116.

7.2 Rationale behind the rule and the Commission's impact assessment

The Commission concludes that the order to execution ratio rule can achieve the same objectives as imposing a minimum resting time (ie, it would stop high-frequency traders and algorithmic traders from testing the depth of order books by submitting and cancelling orders in very quick succession, thereby putting less stress on the IT systems of market operators and reducing the risk of systemic failures), but that the disadvantages would be less severe:

> This measure would in all likelihood, provided that the ratio is suitably calibrated, only affect the high frequency traders or algorithmic trading activity it is targeted at. Market liquidity and efficiency and the quality of price discovery should not be adversely affected. Assuming that the ratio and the system of penalties is effectively calibrated then risks would be effectively addressed while minimising the adverse effect on spreads. Market operators would be best placed to calibrate the optimal approach that fits for the particular market concerned.[77]

7.3 General impact

Although discussions with market participants have indicated that different types of strategy result in different levels of cancellation to execution ratios, there are a few general considerations that may dominate the impact of minimum execution ratios. It should also be noted that high-frequency traders can trade profitably only if they trade—that is, unless some orders execute, traders cannot continue in business. (It should also be noted that, as a group, high-frequency traders can only remain in business if they trade with non-high-frequency traders—ie, they provide a service for which non-high-frequency traders will pay.)

The first consideration is that high-frequency AT will tend to use probabilistic analysis and hence probabilistic trading strategies. Therefore, at the margin, as something like a minimum ratio rule bites, the response is likely to be to eliminate the order sequences that have a lower probability of leading to some order execution (ie, an actual trade).

The second consideration is that orders will currently tend to be cancelled when they have a probability, rather than a certainty, of executing at the wrong time. At present an order can be cancelled without any real costs when it has a low probability of executing at the right time, even if the risk of executing at the wrong time is very low. In the presence of a minimum execution ratio requirement the probability of executing at the wrong time would tend to dominate the cancellation decision, rather than any consideration of the low probability of executing at the right time.

Third, if there is a wide definition of AT which will include most trading that occurs, a minimum ratio requirement will have the unintended consequence of favouring traders who are using both HFT strategies as well as executing strategies designed to produce significant changes in net positions. Small boutique high-frequency traders will be at a disadvantage to large, multi-strategy traders who are both using HFT (for example, as a proprietary trading strategy) and executing a significant amount of agency trades for, say, mutual funds.

Finally, it may also be possible for high-frequency traders to 'create' executing trades if required. Although such an approach would incur the trading costs of using the relevant venue,

[77] European Commission (2011), 'Commission staff working paper, Impact Assessement accompanying the document: Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p.129

it could be carried out without any market impact or loss across the spread because the trader would be on both sides of the trade. For example, if a high-frequency trader could establish that a tick inside the touch was empty it could send both a buy and sell order at that price to the venue and get credit for two transactions. Alternatively, if a high-frequency trader could predict accurately that it was at the head of the queue at the touch it could trade with itself (at a high probability) by sending a matching marketable order at that precise time. It might even be possible for two traders to coordinate if trading with oneself in these circumstances was discouraged.

As a result of these considerations it should be possible for high-frequency traders to modify their order flow to venues in ways that would meet the required ratio, but without having a really significant impact on the orders that do, actually, execute. However, at the margin, a minimum ratio requirement could be expected to have some impact on behaviour, with a reduction in the placing of orders that have the highest probability of executing at the wrong time if left, or the lowest probability of actually being wanted at all. These order types will include:

- positioning orders placed a long way from the touch;
- pinging orders with a low probability of executing on arrival;
- multiple orders to different venues, only one of which is required to execute.

As a result of these considerations, and because high-frequency traders are likely to use probabilistic approaches to meet a requirement of this sort, any impact on HFT will be at the margins (unless the ratio is set very tightly).

Although, it can be seen that the imposition of minimum ratios will have some impact at the margin, the Commission's implicit assumption that there is a bright line in terms of ratios when only unwanted trading would be affected would not appear to hold. There are likely to be significant numbers of orders that a high-frequency trader could decide not to send because the probability of execution is very low, without having a serious impact on sending orders that do have a reasonable probability of execution. (It is worth reiterating that high-frequency traders generally only make money when they do execute.) As ratios were tightened behaviour would be modified, but would generally not stop.

8 Strengthening organisational requirements (circuit breakers)

8.1 The proposed rule

The MiFID review proposals contain three types of organisational requirements: authorisation for high-frequency traders that are members of a regulated market or an MTF; stricter risk controls for firms engaged in automated high-frequency trading; and mandatory adoption of system stress-testing and circuit breakers by market operators. The first two of these initiatives are purely changes in administrative burden on a sub-class of investment firms, with low likelihood of indirect costs or material effects on the market structure. In contrast, the reinforced organisational requirements for market operators aimed at reducing the risk of disorderly trading (eg, circuit breakers) will affect how markets operate for all market participants and have a much larger scope for indirect consequences. This section will therefore focus on this subset of the proposals.

The final version of the Commission's proposals describes the new requirement as follows:

> Member States shall require a regulated market to have in place effective systems, procedures and arrangements to reject orders that exceed pre-determined volume and price thresholds or are clearly erroneous and to be able to temporarily halt trading if there is a significant price movement in a financial instrument on that market or a related market during a short period and, in exceptional cases, to be able to cancel, vary or correct any transaction.
>
> The Commission shall be empowered to adopt delegated acts ... to set out conditions under which trading should be halted if there is a significant price movement in a financial instrument on that market or a related market during a short period.[78]

The new requirements do not, at least at this stage, prescribe specific design for circuit breakers or price limits. The intention to empower the Commission to adopt delegated acts, however, suggests that further detailed guidance on specific features of the new operational requirements may be forthcoming. At this stage it is not clear how much discretion the final rules will leave to national authorities or individual exchanges.

Another noteworthy development is that, although the main focus of the operational requirements continues to be on automated trading halts—circuit breakers—the latest version of the policy document also refers to rejection of orders that exceed pre-determined price thresholds, a measure more in line with daily price limits or the 'limit up–limit down' proposals currently being discussed in the USA.

8.2 Rationale behind the rule and the Commission's impact assessment

The Commission's main rationale behind the new operational requirements is mitigating the risks of disorderly trading and order execution away from fundamentals that may arise as a result of AT or other unexpected shocks to the markets.[79] The impact assessment also cites the similarity of the proposed operational requirements to the regulatory measures considered in the USA as a further supporting factor. Although the MiFID II impact assessment does not elaborate on these statements, the basic rationale for introducing measures to curb volatility and prevent trading errors appears intuitive.

In general, market crashes and disorderly trading may result in costs to certain traders and impair market confidence. Using circuit breakers to prevent crashes and reduce the risk of significant market volatility would therefore be expected to be beneficial. Most, if not all, traders interviewed by Oxera also confirmed that, from their perspective, circuit breakers were an effective way to control abnormal market volatility.

One caveat, however, is the extent to which envisaged benefits of the new rules will be incremental to existing practice. Having circuit breakers is not a regulatory requirement in Europe at present, but responses to CESR/ESMA's call for evidence on market microstructure

[78] European Commission (2011), 'Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p. 116.

[79] European Commission (2011), 'Commission staff working paper, Impact Assessement accompanying the document: Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments', p.127

indicate that almost all European trading venues already have such controls in place.[80] In fact, even before the new MiFID proposals, some academic commentators cited the widespread use of circuit breakers in Europe to argue that the flash crash of the magnitude observed in the USA could not occur in European markets.[81]

The final version of the MiFID II impact assessment argues that the incremental costs to exchanges of complying with the new circuit breaker and stress-testing requirements would be minimal; in effect, the new rules would enshrine existing practice. For example, the estimated ongoing costs of the proposed operational requirements for the exchanges are nil. This appears to be inconsistent with the large (unquantified) benefits attributed to the new rules, as well as with the observation that some European exchanges do not have the required measures in place.

8.3 Effects on markets overall

A specific advantage of circuit breakers, compared with the other rules proposed by the Commission, is that it is an ex post rather than ex ante measure. This means that it attempts to control volatility when it occurs, rather than restricting traders' behaviour ex ante with the risk of not only banning detrimental trading strategies or behaviour but also those behaviours that can have positive effects. These insights are also supported by Oxera's interviews with the market participants.

Despite the intuitive attractiveness of having measures specifically designed to curb volatility, the empirical evidence on the effects of circuit breakers and price limits on the markets is inconclusive. It is not clear, however, to what extent the findings in the academic literature can be applied to the new MiFID proposals, as market studies on this subject date back to over five years ago, are often focused on smaller regional exchanges, and, therefore, tend to explore the effects of relatively simple volatility control mechanisms in trading environments before the advent of large-scale AT. Nonetheless, academic research provides interesting insights into the principles of how volatility controls affect the market.

The literature discusses, broadly, two categories of the operational volatility controls that can be implemented by trading venues: price limits and circuit breakers (market-wide or, more rarely, stock-specific).

- Price limits are a widespread method of volatility controls in stock and futures markets, which limit the daily price movement of an instrument to pre-defined bands around its previous day's settlement price.[82] In addition to preventing panic-driven price volatility and excluding destabilising trading activities, the intended benefits of price limits also include limiting daily liability and portfolio adjustment needs of market participants and allowing brokers the time to consult clients when markets are turbulent. These measures have, however, also been criticised for delaying price discovery, interfering with potentially rational, though volatile, trading activity and causing volatility to spill over to next day by preventing immediate price

[80] ESMA (2011), op. cit, p. 76.

[81] See, for example, Gomber et al. (2011), op. cit.

[82] Kim, Y.H. and Yang, J.J. (2004), 'What makes circuit breakers attractive to financial markets? A survey', *Financial Markets, Institutions & Instruments*, **13**:3, pp. 109–46.

corrections.[83] The substantial volume of empirical research about the effects of price limits on the futures and stock markets generally provides mixed and inconclusive results; the main consistent findings, however, are that in stock markets fixed daily price limits delay price discovery and do not have a significant effect on volatility (or increase it). Overall, a comprehensive literature review of price limits by Kim and Yang (2004) concludes that fixed daily price limits appear to be undesirable for stock markets. The authors conclude that more evidence is needed to determine the effects of these measures in futures markets.[84]

- The second class of measures, circuit breakers operate by automatically triggering a temporary trading halt on a security or the market as a whole when price volatility exceeds a predefined threshold. As with price limits, academic literature offers conflicting findings on the effects of circuit breakers. There is some theoretical and empirical evidence that trading halts reduce short-term volatility, provide an opportunity for reassessing available information and reviewing orders, and moderate noise-generated panic, as envisaged in the MiFID proposals.[85] Other empirical studies, however, have shown that trading halts are at best ineffective and potentially may have the detrimental effects of delaying price discovery and increasing volatility after markets re-open.[86] Furthermore, academic literature suggests that if the trigger mechanisms are known to the market, circuit breakers can sometimes have a 'magnet effect'—increasing volatility and exacerbating price movements when traders see that the price is close to the trigger and attempt to execute their trades before the imminent trading halt.[87]

The difficulty with using the empirical results described above to assess the effect of the current operational requirements under MiFID II is that most of the existing academic literature explores old designs of price limits and circuit breakers, which use fixed pre-defined bands to constrain the size of the deviation of the instrument prices from their previous day's closing price. The findings may, therefore, not apply to the modern dynamic circuit breakers used on most large European exchanges,[88] which use the recent instrument prices to constantly re-calculate the trigger price bands throughout the day. Furthermore, in addition to dynamic triggers many modern circuit breakers do not simply suspend trading on a stock once triggered, but implement shorter, two- to five-minute, trading pauses, during which information flows in the market are allowed to continue by the means of an ongoing call auction.

Few empirical studies have been carried out on these mechanisms to date, but existing empirical evidence on the new dynamic circuit breakers that trigger a short-lived pause

[83] See, for example Kim, K.A. and Sweeney, R.J. (2001),'The Effects of Price Limits on Information Revelation: Theory and Evidence', unpublished working paper, p. 115.

[84] Kim, Y.H. and Yang, J.J. (2004), 'What Makes Circuit Breakers Attractive to Financial Markets?: A Survey', Financial Markets, Institutions & Instruments, Vol. 13, No. 3, pp. 109-146, August .

[85] See Lee, W., Park, J.W. and Jordan, J.J. (2009), 'Trading Halts and Information Asymmetry', *Working Paper*, p. 4, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1537367.

[86] Goldstein, M.A. and Kavajecz, K.A. (2000), 'Eighths, Sixteenths and Market Depth: Changes in Tick Size and Liquidity Provision on the NYSE', *Journal of Financial Economics*, **56**:1, pp. 125–49; McMillan, H. and Overdahl, J. (1998), 'Another Day, Another Collar: An Evaluation of the Effects of NYSE Rule 80A on Trading Costs and Intermarket Arbitrage', *The Journal of Business*, **71**:1, pp. 27–53; and Veld-Merkoulova, Y. (2003), 'Price Limits in Futures Markets: Effects on the Price Discovery Process and Volatility', *International Review of Financial Analysis*, **12**:3, pp. 311–28.

[87] The theoretical argument is presented in Subrahmanyam, A. (1994), 'Circuit Breakers and Market Volatility: A Theoretical Perspective', *The Journal of Finance*, **49:**1, pp. 237–54. Empirical evidence on the existence of this 'magnet effect' in practice is mixed (see Kim, Y.H. and Yang, J.J. (2004), 'What Makes Circuit Breakers Attractive to Financial Markets?: A Survey', Financial Markets, Institutions & Instruments, Vol. 13, No. 3, pp. 109-146, August .

[88] See http://www.londonstockexchange.com/about-the-exchange/regulatory/lsegresponsetoesmaconsultationonsystemsandcontrols.pdf and http://www.world-exchanges.org/files/statistics/excel/WFE%20Circuit%20Breakers%20Report%20%28R%C3%A9par%C3%A9%29.pdf.

accompanied by a call auction on the Bolsa de Madrid suggests that this volatility control mechanism has moderated volatility and adverse selection costs. The trading patterns after these circuit breakers are triggered have been found to be consistent with curbing investor over-reaction rather than preventing efficient price adjustments.[89]

Overall, academic studies and Oxera's discussions with a wide range of market participants suggest that circuit breakers, especially the new sophisticated stock-specific methods, are an effective way to curb market panics and potential disorderly trading. Naturally, the new proposals for mandatory circuit breakers or price limits are not without costs, since introducing or modifying volatility controls is likely to require additional system investments by the trading venues. Nonetheless, due to the widespread use of circuit breakers in Europe, it appears that the new operational requirements will necessitate a relatively small departure from current practice, and should therefore have no, or little, incremental costs and benefits. The magnitude of these effects will, however, significantly depend on the specific design requirements for the new measures that may be articulated in subsequent legislation.

None of the HFT strategies and trading sequences identified in discussions with the market participants rely on abnormal price movements for profitability, so introduction of these proposals is unlikely to affect the traders' relative preferences over high-frequency strategies. A possible exception may arise if different trading venues adopt fixed price limits that diverge significantly, and at the margin, the price limit on one exchange constrains a potentially profitable cross-venue arbitrage trade by a high-frequency trader. The likelihood of this situation arising in practice appears low, so it is unlikely to constitute a material cost (or benefit).

Clearly, one of the main purposes of the MiFID proposals on organisational design is to curtail the potential detrimental effects on the market from market volatility and adverse price spirals that can potentially arise from algorithm interactions. If appropriately implemented, these measures should benefit high-frequency traders as they are most directly exposed to potential losses from erroneous trades executed by malfunctioning algorithms at times of high volatility. Furthermore, stress-testing requirements to ensure that market infrastructures are sufficiently robust to deal with spikes in trading activity also appears beneficial to high-frequency traders, which rely crucially on execution speed and are more likely to become unable to trade profitably if their orders are delayed due to the exchange being overloaded. Given increased market resilience to high order volumes, high-frequency traders may be less likely to withdraw from the market and cease supplying liquidity in market conditions where there is a sudden surge in messaging activity, as was the case in the US flash crash.

Some commentators have cautioned against adopting highly sophisticated automated volatility controls, such as dynamic circuit breakers or price limits, because doing so introduces additional complexity into a system that is already poorly understood due to interactions between trading algorithms. [90] In this environment, there may be unpredictable unintended consequences of the new rules, such as, for example, a rise in volatility within the allowable price bands. Such potential changes in the behaviour of complex market systems as a result of the new rules will need to be carefully researched and monitored.

[89] Abad, D. and Pascual, R. (2008), 'Switching to a Temporary Call Auction in Times of High Uncertainty', Working Paper, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=901425.

[90] See, for example, Cliff, D. (2011), 'The Global Financial Markets: an Ultra-Large-Scale Systems Perspective', *Foresight Driver Review*, DR4.

8.4 Coordination of volatility interruptions across Europe

It is unclear what degree of harmonisation and coordination in volatility interruption mechanisms is envisaged in the final rules. Although the current text of the Commission's proposals requests that the Commission be able to make proposals on specifics of mechanism design, it is difficult to say how much discretion will be left to individual exchanges over specific features of their circuit breakers and price limits under the final rules. Furthermore, the new operational requirements must allow trading venues 'to temporarily halt trading if there is a significant price movement in a financial instrument on that market or *a related market*'.[91] The proposals do not clarify whether this implies that trading halts for a specific instrument need to be automatically triggered on all venues if a trigger is breached on one of the exchanges, or even that trading suspensions need to be linked for a derivative and the underlying security.

It is not clear whether there is a need for coordination. Coordinated suspensions may help to avoid triggers, and may help trader confusion and uncertainty about the continuation of trading (and possibly amplified 'magnet effect' described above) that could otherwise arise when trading venues adopt different triggers and trading suspensions are triggered one by one. On the other hand, coordinated halts could require substantial investment in the European market infrastructure to implement formal links in price feeds and trading instructions among the currently fragmented trading venues. Moreover, universal trading suspension may be inappropriate and disruptive in cases where a large price move on one of the venues is caused by a 'fat finger' error or a technical fault. Generally, if circuit breakers are correctly calibrated, as required in MiFID II, automatic coordination should not be necessary—circuit breakers would trigger in each venue when it was appropriate for curbing volatility for each exchange, and trading suspension would be, in effect, coordinated if shock to the security was genuinely systematic.

The effects of the coordination and harmonisation of volatility controls will need to be considered further when more detail on the MiFID II proposals in this topic becomes available.

8.5 Alternative implementations

An alternative way to address excess volatility is to place dynamic limits on price variation in submitted orders, as is currently proposed in the USA under the 'limit up–limit down' rules.[92] Under this approach orders outside the defined volatility bands are rejected by the trading venue, rather than instigating a trading halt if an executed price is outside the limit. Trading in a security is only briefly suspended if quoted prices remain outside the limits for a defined period of time.

This approach has the benefit of enabling continuous trading, allowing traders to react to all new information as it arrives. Because it is rogue orders that are stopped from being available, trading between non-rogue orders can continue. However, if the rapid price change is a result of a real change in fundamentals (ie, not rogue orders), a limit on the price variation in orders that can be submitted may have the same effect as a trading halt, since there will be no demand for trading within the permissible bands. The main downsides to this alternative include: larger incremental compliance costs of European trading venues implementing a new,

[91] European Commission (2011), 'Proposal for a Directive of the European Parliament and of the Council on markets in financial instruments repealing Directive 2004/39/EC of the European Parliament and the Council', European Commission, October 20th, para 51.

[92] SEC (2011), 'SEC Announces Filing of Limit Up-Limit Down Proposal to Address Extraordinary Market Volatility', pp. 2011–84.

and currently largely unused, volatility control; higher complexity for the market participants; and, potentially, larger trade-processing burden on market infrastructure.

It is not clear at this stage to what extent volatility control mechanisms based on dynamic price limits are substitutable for circuit breakers from the point of view of complying with the current MiFID II requirements.

9 The impact of further increases in speed on the operation of the rules that impinge on trading sequences

As computers get faster, and as the distance between traders' computers and trading venues' computers potentially narrows, the result will be a reduction in the time delay between:

- the information arriving at a trading venue;
- the knowledge of that information being received by the traders;
- the analysis of that information;
- the decision to send a new order or to cancel an existing order and the order/cancellation then being incorporated into the new market description at the trading venue.

The time delay will not, however, get to zero and, perhaps more importantly, the sequencing of the information flows and subsequent updating of the market position, remains the same. No matter how fast high-frequency traders get, it will not be possible for the following to happen, unless the price time rules for trading venue operation are breached.

- The high-frequency trader will not be able to see an order on a trading venue and react so quickly as to get his order in front of the new observed order.
- The high-frequency trader will not be able to see a price change in one venue and make changes in another venue before the price change happens.
- A non-high-frequency trader order that arrives at a trading venue before a high-frequency trader's order, of the same type and for the same security, will maintain time preference over the high-frequency trader's order no matter how fast that trader can react.

As the reaction time of the system reduces, the difference between the knowledge of the state of the system by agents (ie, high-frequency traders) and the actual state of the system reduces. The number of changes in the state of the system that have occurred in the time taken for the high-frequency trader to receive, analyse and act on that description will reduce, at least in so far as orders/cancellations not triggered by HFT considerations are concerned. As speeds increase the high-frequency trader can be slightly more certain that what they are observing and acting on is the accurate description of what the state of the market actually is. Complete certainty is not possible, because there will always be some delay between new information arriving at the trading venue and the high-frequency trader's observation of that change.

Hence, absolute speed does improve slightly the ability of the trader to describe correctly how the market will look at the time any order or cancellation sent by the trader arrives at the trading venue. However, in relation to absolute speed, the improvement can be seen as eliminating the impact of changes that would occur to the system within the reduction in the round-trip latency time. This change will be able to affect decisions that would have been made (eg, send a buy order) but are now changed because a certain bit of information has arrived at the trading venue, and been processed, and is now available to traders *slightly* quicker.

The time window within which this change has to occur is the change in the round-trip latency. With latency currently reported in the 250-microsecond range (ie, ¼ of a millisecond) the probability of an event occurring within a time window representing a change in this latency would be expected to be quite low. In relation to a security with an execution of once every second, and a change in latency of, say, 50 microseconds, the probability of the information arriving at the trader within this window is low—0.005%. This information will only make a difference if it changes the subsequent behaviour of the trader, which is unlikely to be in every case. Clearly, with significantly more orders than executions the knowledge of new orders (or cancellations) in relation to a security will more frequently arrive in this window, but in many cases that information will not change the behaviour of the trader because it does not change the predictions of the algorithms.

These general considerations suggest that the change in absolute speed may have relatively little impact on HFT. However, given the limited availability of profitable trading opportunities using short-term price predictions to the extent that high-frequency traders are in competition with each other for these profits, *relative* speed may be more important.

If relative speed is the more significant characteristic, then those traders wishing to continue to operate in the high-frequency part of the trading market space will need to remain ‘competitive’ with others attempting to use the same techniques. Competition is likely to drive a desire for higher speeds (and more predictive accuracy in the algorithms in general), which may increase the costs of remaining a successful high-frequency trader. However, because high-frequency traders (as a group) must trade with non-high-frequency traders in order to remain profitable, it is the impact on the latter group that should be considered. Here, the primary consideration is likely to be absolute speed, not relative speed, because non-high-frequency traders are already working on reaction times considerably in excess of those of the high-frequency trader.

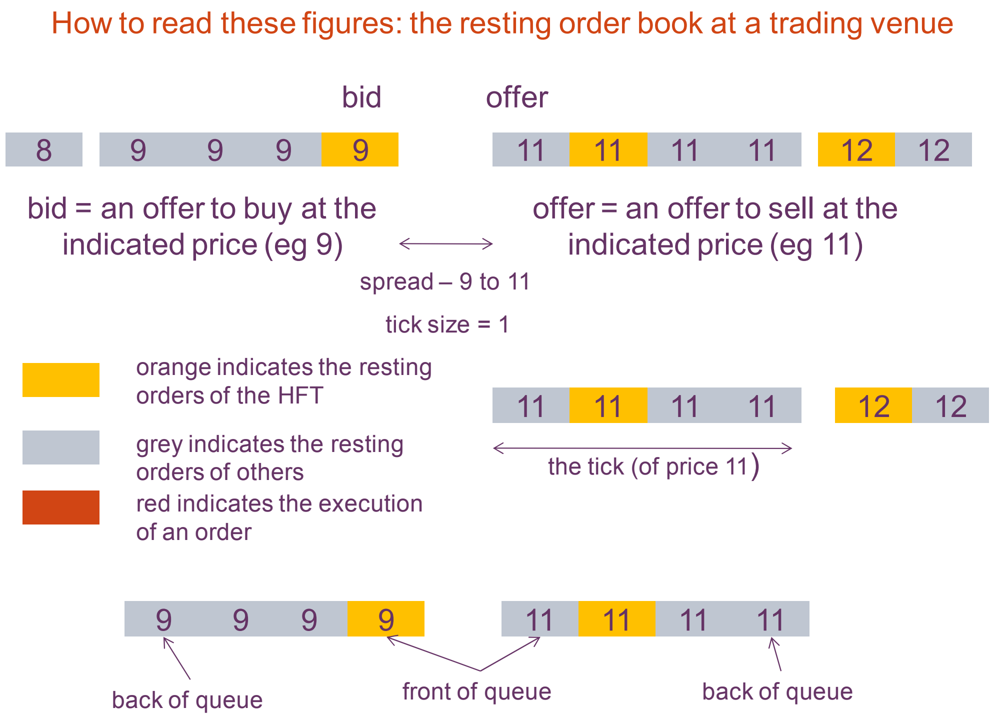
A1 Annex

A1.1 Introduction

This annex sets out in a (very) simplified form the generic trading sequences used by high-frequency traders. The sequences set out in these figures are designed to show the underlying logic behind the trading sequences, but are not a realistic portrayal of any actual order book or actual trading sequence. The source for all of these figures is Oxera.

Figure A1.1 below shows the conventions used in the figures. In particular, the conventions used indicate that orders to the order book will be added at the back of the relevant queue—which is the left side of the tick on the bid side, and the right side of the tick on the offer side. Bids move towards the spread as the resting orders are executed at the front of the queue, or cancelled (at any position in the queue).

Figure A1.1 Conventions used in the figures in this Annex


How to read these figures: the resting order book at a trading venue
bid
offer
8
9 9 9 9
11 11 11 11
12 12
bid = an offer to buy at the indicated price (eg 9)
offer = an offer to sell at the indicated price (eg 11)
spread – 9 to 11
tick size = 1
orange indicates the resting orders of the HFT
grey indicates the resting orders of others
red indicates the execution of an order
11 11 11 11
12 12
the tick (of price 11)
9 9 9 9
11 11 11 11
back of queue
front of queue
back of queue

A1.2 Arbitrage

Figure A1.2 shows the sequence for the simplest arbitrage strategy. Two aggressive orders are used once the pricing anomaly has been identified. To execute a successful trading sequence of this sort requires a relatively large pricing anomaly to be present.

Figure A1.2 Simple arbitrage, aggressive orders

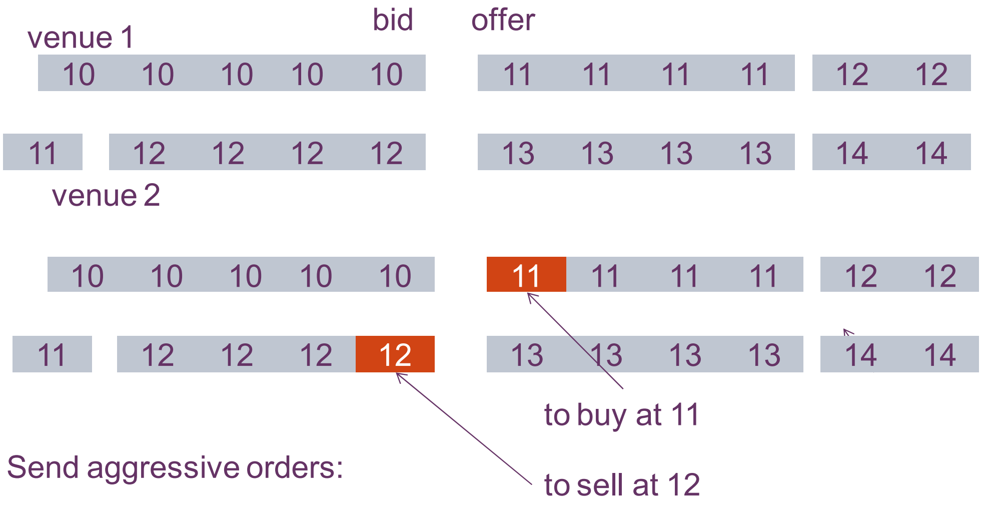
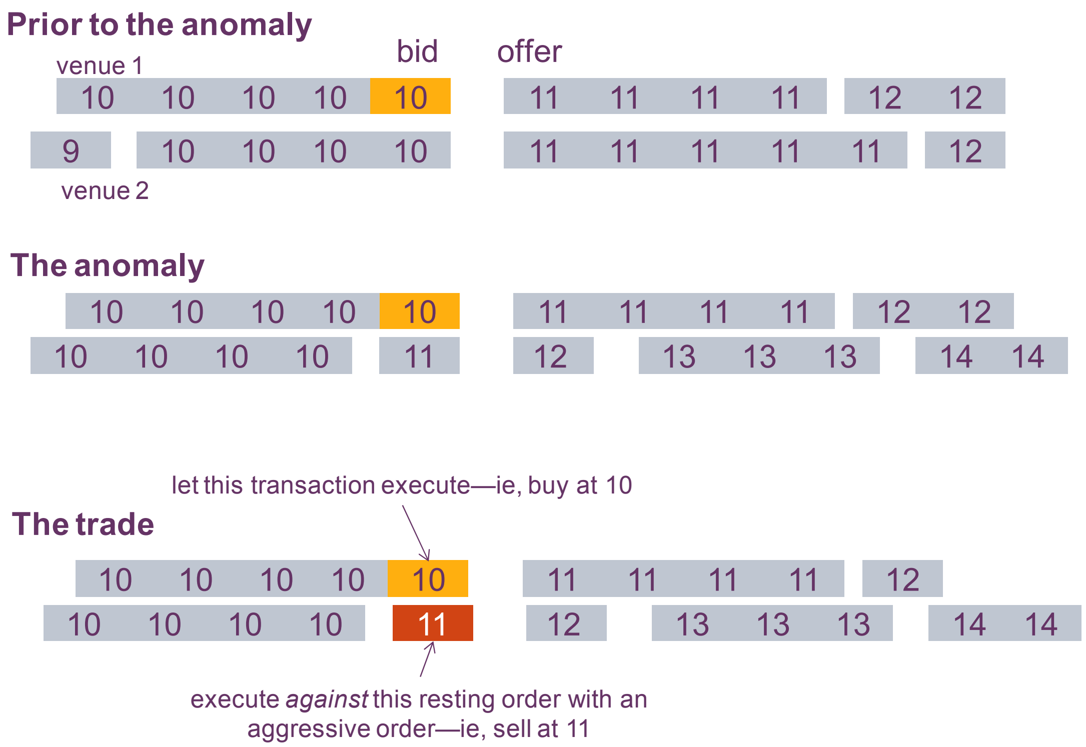


Figure A1.3 illustrates the use of a combination of an aggressive and a passive order to take advantage of the same pricing anomaly.

Figure A1.3 Simple arbitrage, one passive and one aggressive order

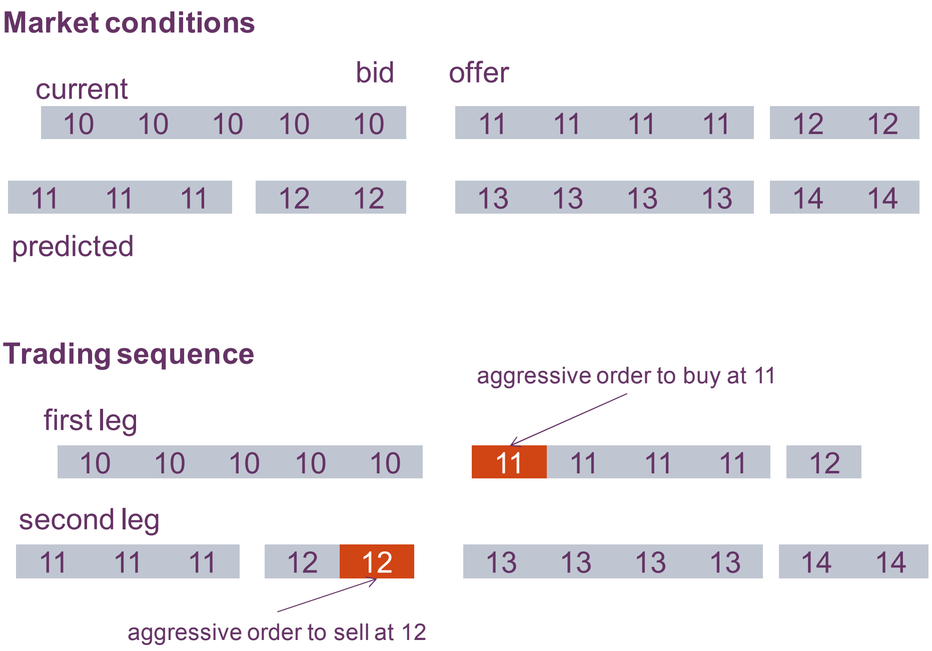


Although it is possible to have resting orders in both venues and to execute those orders when there is a pricing anomaly, the trading sequence will include profiting from the spread. This sequence therefore shares many characteristics with market-making or directional strategies.

A1.3 Directional

Figure A1.4 shows the trading sequence for the simplest directional strategy when a prediction has been made that the price of a security will change. The simplest sequence relies on two aggressive orders for completion, and, as a result, requires a relatively large price change to occur in order to be profitable. (In this example, the midpoint of the spread has to move from 10.5 to 12.5.)

Figure A1.4 Simplest directional strategy



Using passive orders to accomplish this sequence reduces the size of the price movements that are required to undertake a profitable sequence. Figure A1.5 sets out this sequence using one aggressive and one resting order. Figure A1.6 sets out the sequence using two resting orders.

Figure A1.5 Simple directional sequence, one resting and one aggressive order

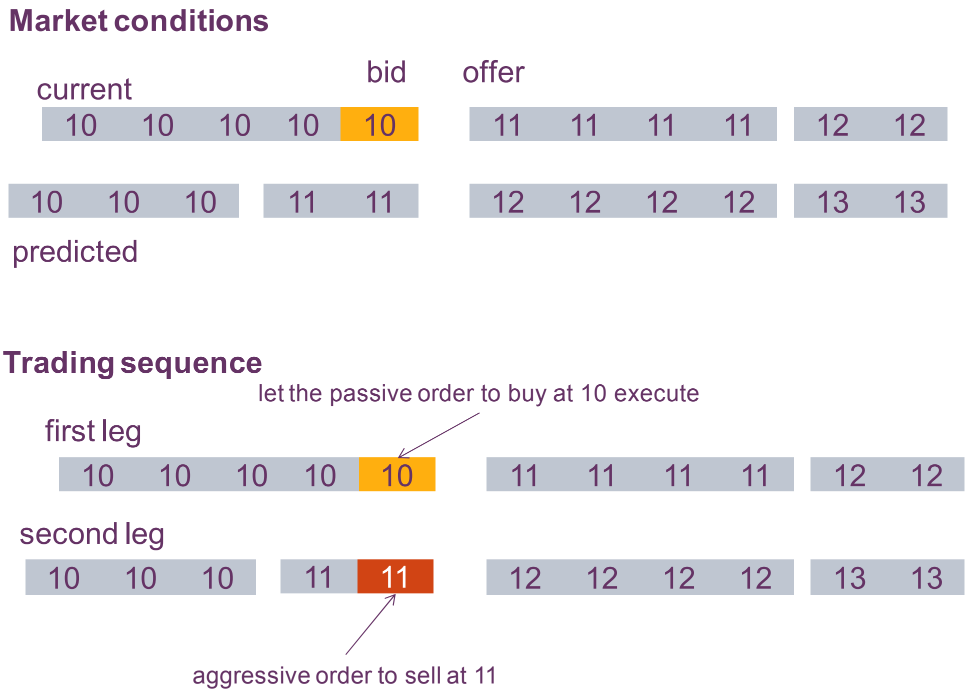
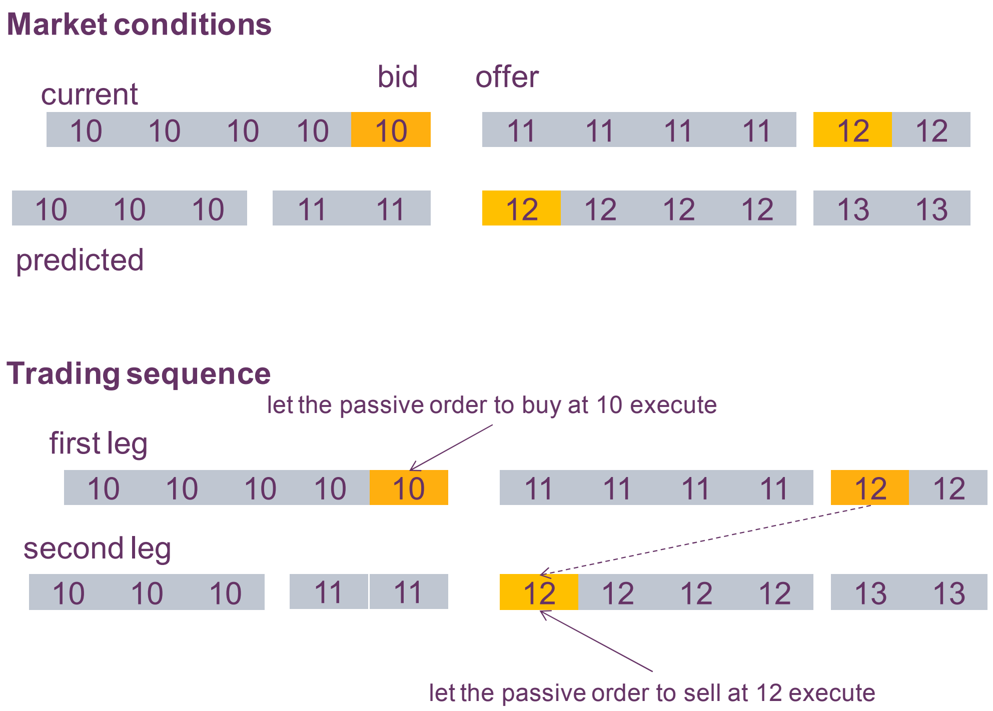


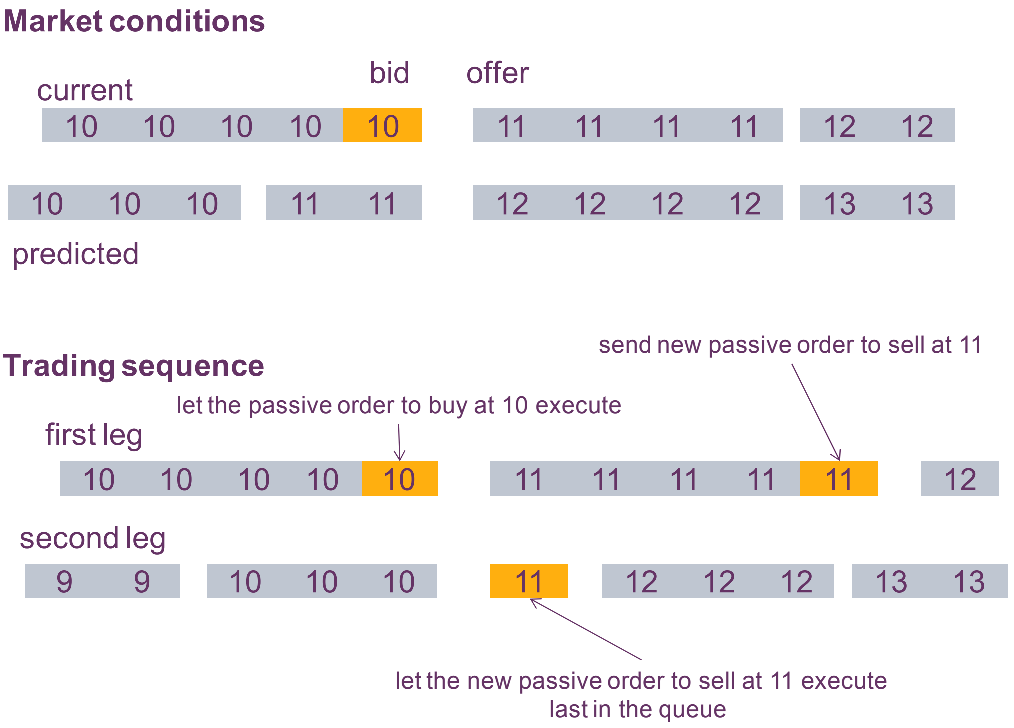
Figure A1.6 Simple directional sequence using two resting orders



The sequences using resting orders can be used with smaller price changes. When using one resting order, any price change that is larger than the transaction costs incurred (eg, exchange fees, clearing charges) is potentially profitable at the margin. When two resting orders are used the sequence takes on the characteristic of a market-making strategy, where there is a benefit to trading even if no price change occurs. With a price change the profitability of the sequence increases (as long as the price change is in the right direction).

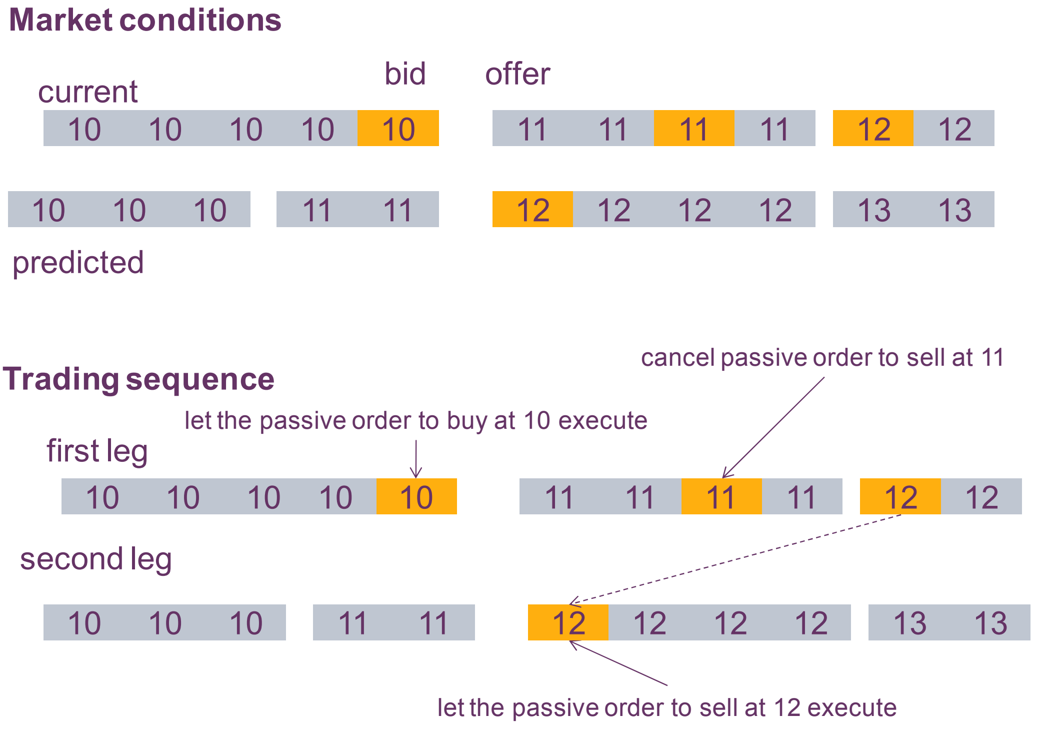
Where a trader does not have a suitable resting order in the right tick (according to the prediction) to provide the second leg of the sequence a suitable order will need to be sent as the first leg executes. Figure A1.7 illustrates the placing of the order at the back of the queue in the tick that will be exhausted if the prediction turns out to be correct.

Figure A1.7 Placing an order at the back of the right queue for the second leg



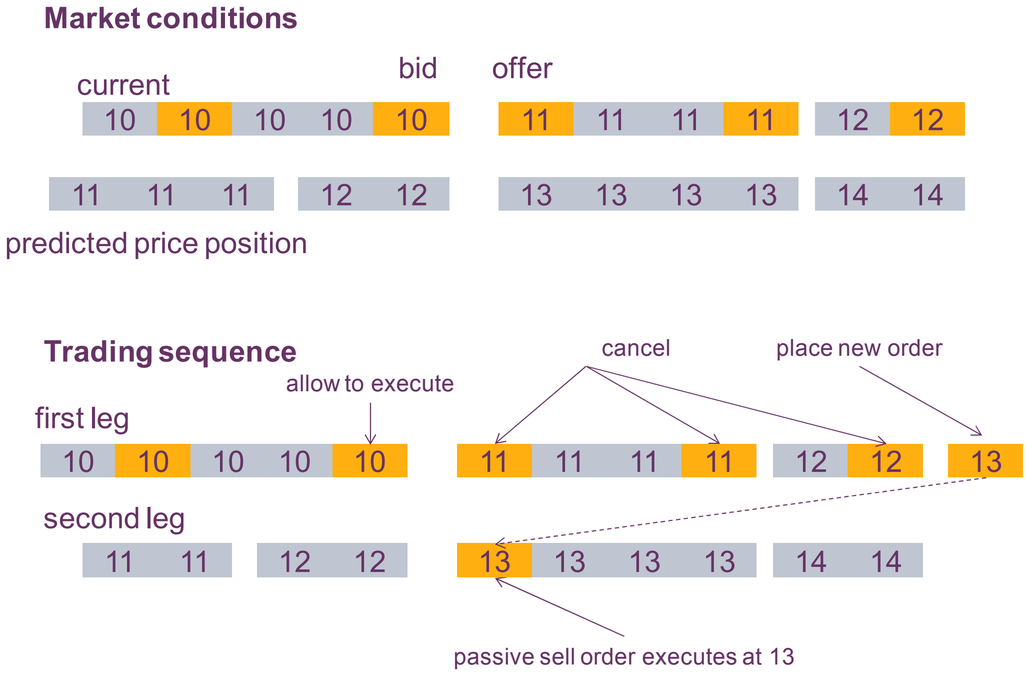
By maintaining resting orders in different ticks, the trader can choose to let the right resting order execute depending on the strength of the predicted price movement. Figure A1.8 depicts the simplest sequence when the price movement would take the price through the first resting order maintained by the trader.

Figure A1.8 Predicted price movement through the first resting order



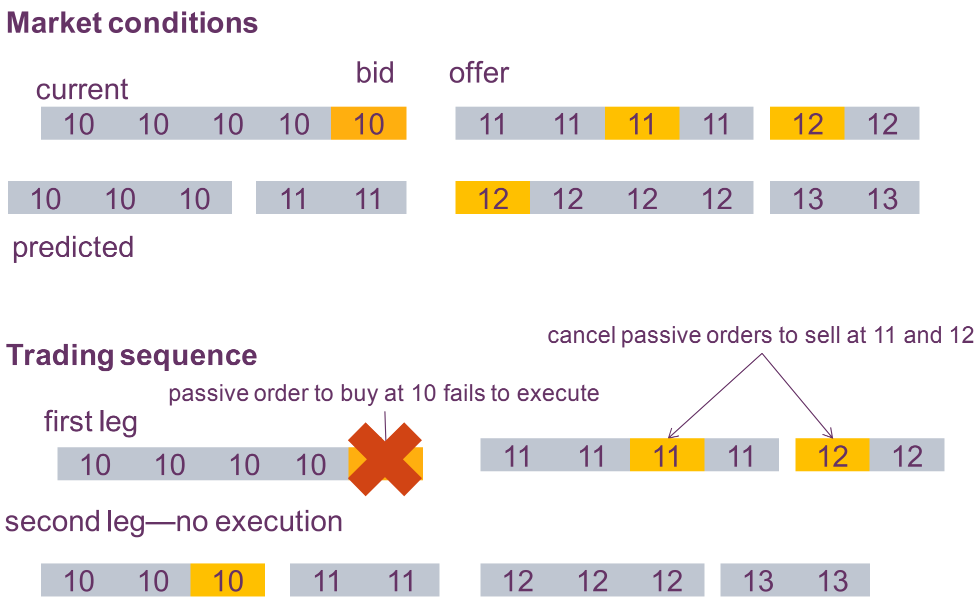
By maintaining resting orders on both sides of the spread, a trader can attempt to use predicted price changes in a way that minimises that their costs. However, the optimal trading sequence to take advantage of any (predicted) price change may not involve executing the trader's existing resting orders if they were just left to execute. Figure A1.9 shows the trading and order sequence needed by a trader when starting from a neutral position (ie, a position from which the trader can use a predicted price change in either direction), and where the predicted price movement is beyond where the trader currently has resting orders. (In practice, traders are likely to have resting orders further into the order book than illustrated here.)

Figure A1.9 Complex sequence from a neutral position



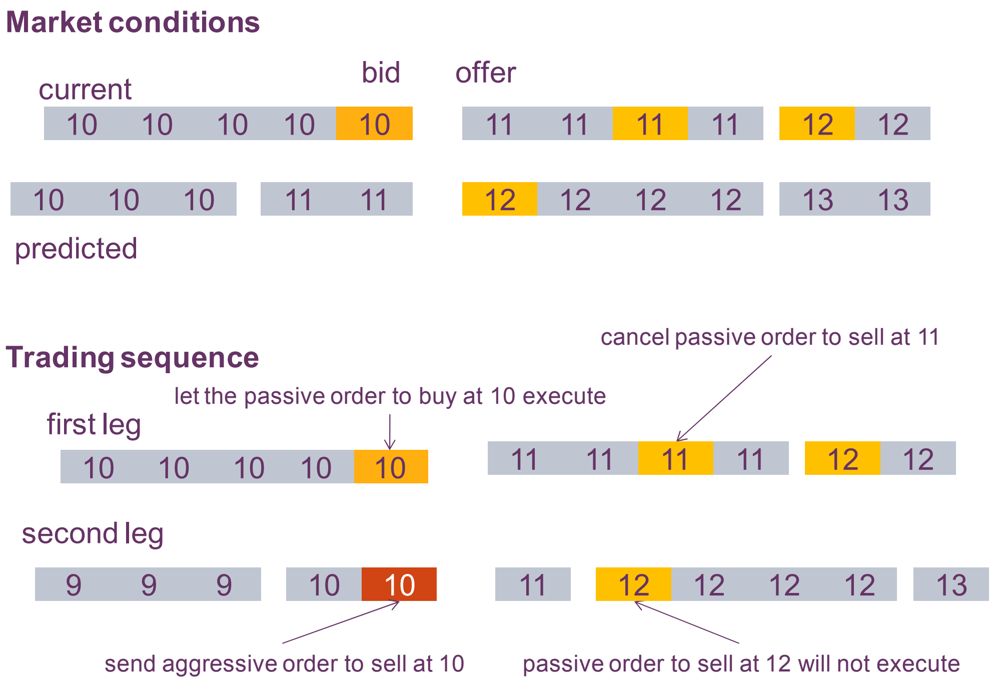
Not all predictions will turn out not to be correct, and it is also possible that the first leg of a sequence will fail to execute before the price movement starts. When this happens, further cancellations are required to avoid the second leg of the sequence executing in the absence of the first leg. Figure A1.10 sets out the sequence when the first leg fails to execute.

Figure A1.10 Failure of the first leg to execute



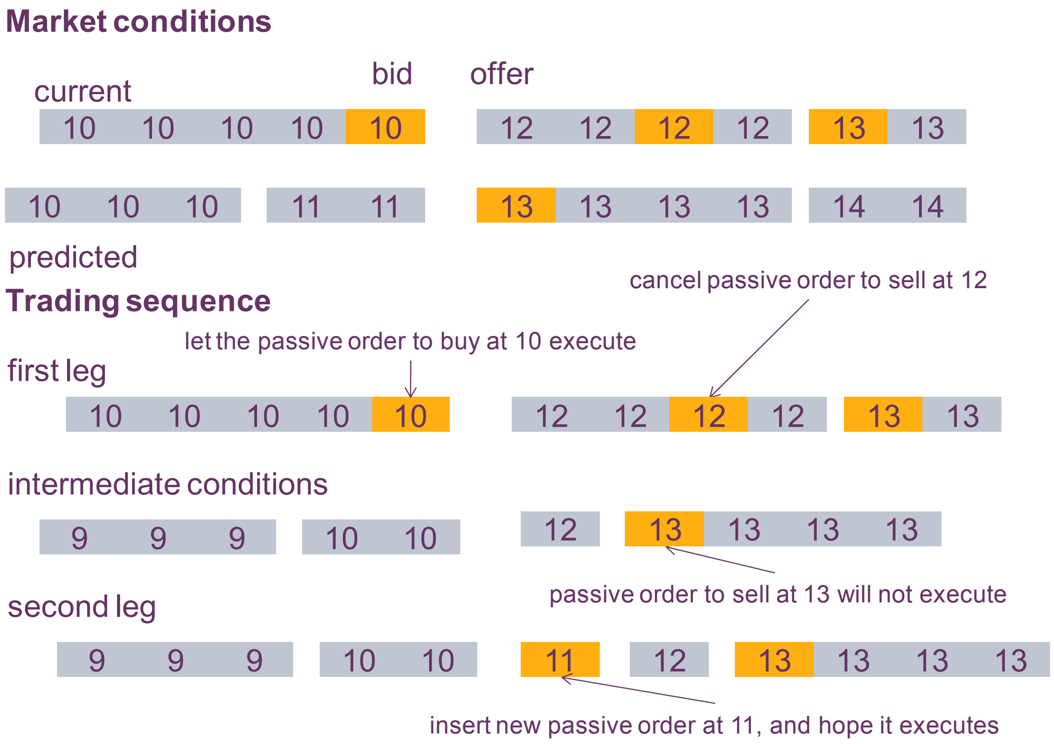
Where the price movement fails to reach the level predicted, the trader runs the risk of the second leg failing. This can be corrected by sending more orders or by using an aggressive order, but the opportunity to profit from the trade may be lost. Figure A1.11 illustrates the sequence using an aggressive order to unwind the first leg (in this case, at the same price).

Figure A1.11 Failure of price change to reach its predicted level



An alternative approach is for the trader to attempt to insert a passive order into the order book inside the spread. This can result in the trader not losing (all of) the spread. Figure A1.12 illustrates this approach.

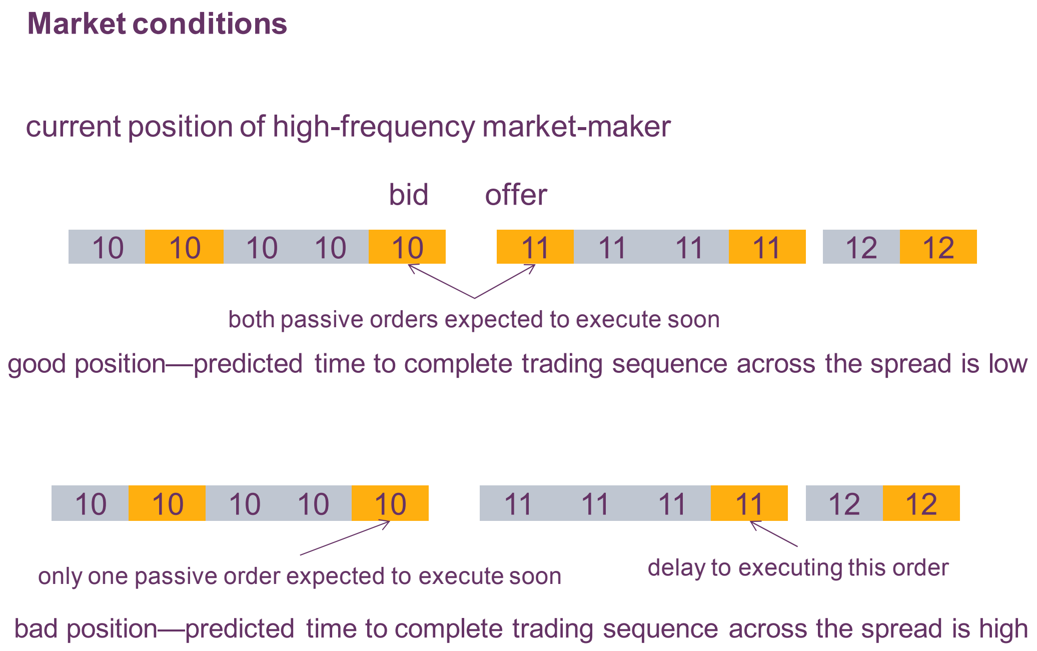
Figure A1.12 Using a resting order inside the spread to recover on the second leg



A1.4 Market making

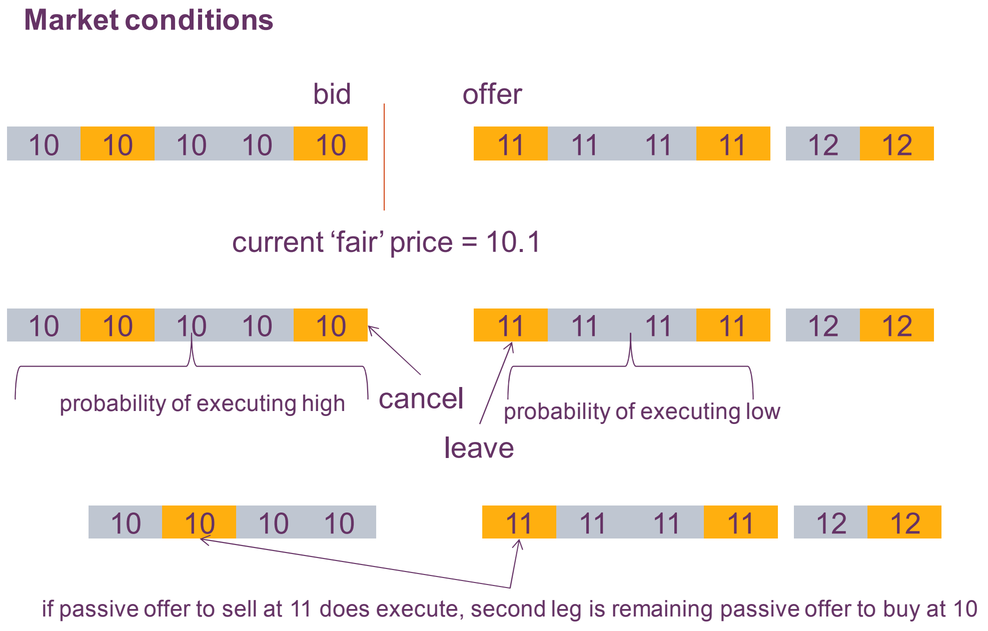
In the absence of a price movement, the high-frequency trader may adopt a pure market-making approach. In order to minimise the probability of an adverse movement in price, and to reduce the use of capital by minimising holding times, the conditions under which a market-making sequence is allowed to take place can be controlled. Figure A1.13 illustrates the positioning of resting orders when a market-making sequence would be lower-risk and high-risk.

Figure A1.13 Minimising timing risks in market making


Market conditions
current position of high-frequency market-maker
bid
offer
10 10 10 10 10
11 11 11 11
12 12
both passive orders expected to execute soon
good position—predicted time to complete trading sequence across the spread is low
10 10 10 10 10
11 11 11 11
12 12
only one passive order expected to execute soon
delay to executing this order
bad position—predicted time to complete trading sequence across the spread is high

Even where a market maker is in a good position in the queues at the touch, the probability of executing on each side may not be the same. If there are more marketable buy orders than sell orders, for example, a resting order to sell that is at the head of the queue is more likely to execute than a resting order to buy that is also at the head of its queue. This can be understood as the fair price for the security being closer to the resting sell offers. Under these circumstances, minimising holding times (which uses capital) may involve cancelling resting orders that would execute on the easy side of the touch, until such time as one of the trader's orders on the difficult side executes. Figure A1.14 illustrates this.

Figure A1.14 Market-making sequence in the presence of order flow imbalance


Market conditions
bid
offer
10 10 10 10 10
11 11 11 11
12 12
current 'fair' price = 10.1
10 10 10 10 10
11 11 11 11
12 12
probability of executing high
cancel
leave
probability of executing low
10 10 10 10
11 11 11 11
12 12
if passive offer to sell at 11 does execute, second leg is remaining passive offer to buy at 10

Because the order flow imbalances are likely to be constantly changing where orders on the easy side are cancelled when they get to the front of the queue, they are also likely to be repositioned at the back of the queue. This is both to ensure that there are resting orders available to execute should an order on the hard side execute, and also to ensure that the trader has orders available when that tick becomes the hard side and the trader wants that execution as the first leg.

The desirability of maintaining orders in the queues so that they are available to execute at the right time, while ensuring that orders do not execute at the wrong time, can result in high cancellation rates compared to execution rates. In practice, the considerations that a trader is likely to take into account are much more complex than those set out above, and the right conditions may be relatively rare. As a result, (very) high cancellation rates can arise from pursuing the trading sequences outlined above.

Foresight
1 Victoria Street
London SW1H 0ET
www.foresight.gov.uk
URN: 12/1080